

Università degli Studi di Trieste

Ph.D. in Applied Data Science & Artificial Intelligence

XXXVIII cycle

Towards End-to-End Explainability in Food Science

Candidate
Leonardo Arrighi

Supervisors
Prof. Sylvio Barbon Junior
Prof. Michele Simonato
Prof. Luca Bortolussi

Summary

In the food industry, Artificial Intelligence (AI) techniques play a fundamental role, especially in product quality analysis, a field that requires processing large volumes of heterogeneous data and handling the complexity inherent in the subjective assessment of quality that is often needed in this sector. In response to the need for reliable AI models, EXplainable Artificial Intelligence (XAI) has emerged, providing explanations of how models make decisions and improving their transparency and interpretability. XAI is therefore relevant in the food industry, since it makes AI technologies more transparent, safer, and applicable even to delicate products such as food.

This doctoral thesis explores the application of XAI methods to the analysis of food product quality, proposing innovative solutions that integrate into an XAI pipeline across the food supply chain. More specifically, tabular data, commonly used in the food industry to describe the physicochemical properties of food products, are analyzed using tree-based ensemble models, which prove to be effective in addressing various food engineering challenges.

Additionally, to enhance the interpretability of these models, a new XAI technique called Decision Predicate Graphs (DPG) is introduced. DPG is a graph structure that captures relationships among features, logical decisions, and model predictions. The effectiveness of DPG is demonstrated in three case studies across three fruit types at different ripeness stages, using physicochemical data to improve AI-based predictions of food quality parameters.

The thesis examines the development of a complete AI workflow for food quality analysis. In this context, XAI is necessary not only for understanding the final model but also for improving data preparation, reducing bias, and boosting performance. For this reason, an extension of DPG is proposed to explain Isolation Forest, a tree-based ensemble technique used to identify potential outliers in the preprocessing stage of an AI workflow. This approach provides a comprehensive view of the decision process in outlier detection, enabling an end-to-end XAI-based pipeline.

Finally, the proposed methods are evaluated in real case studies, assessing their effectiveness in critical operational contexts.

In conclusion, this thesis highlights the effectiveness of XAI methods in improving the analysis of food product quality, making predictions more accurate and reliable, reducing bias, and supporting the adoption of advanced AI technologies to ensure high standards of quality and safety.

Contents

List of Abbreviations

^{1}H NMR Proton Nuclear Magnetic Resonance. 25, 35

ADD Algebraic Decision Diagram. 48, 49, 58–60, 107

AI Artificial Intelligence. 1–6, 8, 11, 17–19, 21–23, 25–27, 29, 30, 32–34, 42, 63–66, 74, 75, 77, 95, 103, 104, 107

AI Act Artificial Intelligence Act. 2, 103

BC Betweenness Centrality. 52, 55–57, 60, 61, 70, 72, 74, 97, 100, 101, 109, 110

BELLATREX Building Explanations through a LocalLy AccuraTe Rule EXtractor. 5, 6, 8, 48, 63, 65, 68–74, 104, 110

CAM Class Activation Mapping. 20, 23, 27–29, 43, 44

CATBoost Categorical Boosting algorithm. 33, 35, 39

CNN Convolutional Neural Network. 12, 17, 18, 20, 26–32, 34–40, 44

DIFFI Depth-based Isolation Forest Feature Importance. 78, 80, 93

DL Deep Learning. 1, 3, 12, 18, 22, 26, 29–32, 35, 36, 38, 39, 41, 65

DNN Deep Neural Network. 33, 39

DPG Decision Predicate Graph. 5–8, 10, 45–47, 49–63, 65, 68–70, 72, 74, 77–85, 87, 88, 90–93, 95–97, 99–101, 104, 105, 107, 109, 110

ELM Extreme Learning Machine. 26, 29, 36

FHB Fusarium Head Blight. 27, 30

FI Feature Importance. 3, 31, 33, 34, 39, 53, 55, 57, 68, 74, 78–81, 93, 109

GAN Generative Adversarial Network. 12, 26

GDPR General Data Protection Regulation. 2, 96, 103

Grad-CAM Gradient-weighted Class Activation Mapping. 20, 23, 27–32, 35–40, 43

HSI Hyperspectral Imaging. 25, 30, 32, 34, 38

iForest Isolation Forest. 3, 5–8, 10, 13, 15–17, 47, 77–83, 85–93, 95–101, 104, 105, 107, 110

IOP-Score Inlier-Outlier Propagation Score. 8, 78, 79, 82–93, 97, 99, 100, 104, 110

IoT Internet of Things. 34, 41, 42, 64, 65, 75

iTree Isolation Tree. 16, 91

kNN k-Nearest Neighbors algorithm. 12, 26, 34, 35

LIME Local Interpretable Model-Agnostic Explanations. 20, 23, 28–31, 33–36, 38, 39, 41–44

LR Linear Regression. 17, 26, 31, 35

LRC Local Reaching Centrality. 53, 55, 57, 61, 70, 72, 74, 97, 100, 101, 109, 110

LRP Layer-wise Relevance Propagation. 20, 23, 30, 32

MDI Mean Decrease in Impurity. 57

ML Machine Learning. 1, 4, 6, 7, 11, 12, 22, 26, 29, 32–35, 37, 39–41, 46, 54, 65, 77, 91, 95–97, 101, 107

MS Maturity Stage. 67, 69–71

NIR Near-InfraRed. 25, 37

NN Neural Network. 12, 26, 37, 41

OOB Out-Of-Bag. 15

PDP Partial Dependence Plots. 20, 23, 34, 35

ResNet Residual Network. 26, 28, 30–32, 36–38

RF Random Forest. 3, 5–8, 12–15, 26, 33–35, 37, 39–41, 46–48, 54–61, 63, 65, 68–74, 79, 96–101, 103, 104, 107, 109, 110

SHAP SHapley Additive exPlanations. 20, 23, 28, 30–37, 39–44, 78–80, 93, 96

SVM Support Vector Machine. 26, 30, 33, 35, 41

TPE Tree-structured Parzen Estimator. 33

UAV Unmanned Aerial Vehicle. 32, 40

VGG Visual Geometry Group. 12, 26, 28, 30, 36, 38

ViT Vision Transformer. 17, 26, 28, 29, 31, 32, 40

XAI EXplainable Artificial Intelligence. 2–4, 6–8, 11, 17–23, 27–40, 42–47, 62, 65, 68, 72, 74, 75, 77–80, 84, 85, 91, 93, 95–97, 103–105, 107, 109

XGBoost eXtreme Gradient Boosting algorithm. 12, 26, 33–35, 37, 39, 103

YOLO You Only Look Once. 12, 26–29, 31

Chapter 1

Introduction

1.1 Motivations and Approach

Artificial Intelligence (AI) is spreading through daily life and industry, reshaping decisions, operations, and quality control. From personal devices and online services to logistics and large-scale manufacturing, algorithmic systems are becoming part of the workflow for both citizens and businesses. Driven by competition and cost pressures, companies adopt AI to deliver more efficient, precise, and cost-effective results, often faster than human workflows or conventional mechanical systems. Across many fields, AI supports quality control, demand forecasting, process optimization, and real-time decision making [23, 121, 203, 266]. Furthermore, the two branches of AI, Machine Learning (ML) and Deep Learning (DL), are increasingly used to extract patterns from complex datasets. By allowing computers to learn directly from data, these techniques support automated decision-making in tasks like classification, forecasting, and data generation.

The food sector mirrors these trends. Along the supply chain from primary production to processing, distribution, and retail, extensive and varied data sets are now routine [159]. Used well, these streams can enable safer processes, tighter specifications, and early detection of deviations. AI also supports food quality control with systematic and fast analysis, and can be deployed to assist non-experts throughout the supply chain. Farmers can rely on remote sensing, drones, and Internet of Things sensors to track soil conditions, weather, and crop health. Harvesters can apply computer vision to detect defects and estimate yield. Retailers can use forecasting and outliers detection to improve traceability and reduce waste. These are only a few areas where AI, and particularly advanced ML and DL techniques, can play a central role.

The rapid adoption of these models raises important questions about transparency. According to [21], transparency is defined as a characteristic of a model that makes it understandable on its own. This transparency is characterized by three properties. First, a human should be able to simulate or think through the model from start to finish. Second, it should be possible to explain each component of the model, including the inputs, parameters, and calculations. Finally, there should be an ability to analyze the mathematical process the model uses to generate its outputs. While these properties do not provide an objective, analytically measurable

definition of model transparency, they do provide a guideline for defining a model as understandable and thoroughly evaluable. Consequently, it is not surprising that AI systems, which are developed through automated processes that result in a big number of interconnected parameters, are often perceived as opaque tools, raising doubts about their reliability [2, 212].

In the food industry, the stakes for precision are high because errors directly compromise food safety, regulatory compliance, and public trust. [21, 223] define trust as the confidence that a model will perform as expected when confronted with a specific problem. Establishing this trust remains a fundamental hurdle because, as previously noted, AI models lack transparency. Their inherent opacity, frequently described as a *black box* (or *opaque box*) nature, presents significant risks. These models can evolve unpredictably over time or inadvertently internalize biases from historical datasets. Consequently, the outputs they produce are often difficult for auditors and industry practitioners to interpret. This lack of clarity creates a dangerous disconnect between high predictive performance and actual operational safety.

In strictly regulated environments such as the food supply chain, systems must be both accurate and fully auditable. Quality and safety protocols must adhere to rigorous legal standards while simultaneously managing the natural variability of ingredients, equipment performance, and environmental conditions. Consequently, the opaque nature of these models prevents rigorous evaluation, rendering them unreliable and unsuitable for high-risk applications such as those in the food sector. To address the opacity problems, the European Commission promotes transparency in AI. It leads to regulation to protect ethics and fundamental rights through two main regulatory frameworks: the General Data Protection Regulation (GDPR) [52] and the Artificial Intelligence Act (AI Act) [74]. The first includes the right to explanation, allowing individuals to obtain meaningful information about the logic used in automated decisions. The second defines a risk-based hierarchy of AI systems, with corresponding transparency, fairness, and accountability requirements.

This doctoral project fits within this industrial and methodological context. ASAC srl, a technological partner for product development in the food sector, sponsored the work and provided sustained oversight, granting access to operational constraints, strategic priorities, and production datasets. The collaboration made clear the need for reliable AI that can be implemented in regulated and risk-sensitive settings. A promising direction is EXplainable Artificial Intelligence (XAI). XAI approaches aim to clarify the complex decision-making of AI models, make their reasoning explicit, and provide evidence that strengthens confidence in both models and data. By opening the so-called opaque box, they enhance auditability, support error analysis, and enable safer deployment [145, 274].

Among the many steps in the food supply chain, food quality analysis is both critical and receptive to innovation. It involves assessing product quality standards, risk factors, and environmental contamination, nutritional value, and the conditions of storage and distribution. The work centers on several sensitive points: rigorous specifications, heterogeneous measurements and data, and a direct link between model outputs and corrective actions, while offering ample opportunities for data-driven improvement. Yet a review of current practice shows a clear gap between the promise of transparent AI and its translation into everyday quality workflows.

Interpretability is widely endorsed in principle, but practical, validated, and easy-to-adapt tools remain scarce in this domain. XAI can help close this gap by providing clear explanations of how models make decisions. It links algorithmic outputs to expert judgment, supports model validation and compliance, and strengthens collaboration between people and machines. XAI is not only a guide for model selection, but also a way to increase the acceptability of a chosen model, understand its outputs, check for bias or procedural errors, and build confidence in online deployment.

It is also important to note that food quality work comes with specific constraints. Models must handle diverse data formats, run with low latency and limited resources, and generalize across product lines and product types. Consequently, explanations should be concise and accessible to process engineers, quality managers, and auditors who are not AI specialists.

Additionally, it is worth noting that many food quality assessment tasks rely on tabular data such as physicochemical measurements, spectral intensities, sensor aggregates, time-aligned process variables, and derived indicators. These data are relatively easy to collect and closely aligned with established quality protocols.

Given the ability to collect tabular data, the possibility of a faster and less computationally intensive training process, and the necessity for clear explanations that we thoroughly discussed, tree-based ensembles such as Random Forest (RF) are often a strong choice. They can rival or even surpass DL architectures in specific setups [158, 171]. Yet they remain underused in modern industrial pipelines, which often favor deep end-to-end models. Tree-based ensembles offer advantages that align well with the food supply chain, including robustness to heterogeneous features, missing data, and outliers; the ability to estimate uncertainty and metrics; strong generalization across scenarios; and a wide range of XAI techniques that are readily available and computationally efficient.

The role of tree-based ensembles extends beyond the final prediction. They are widely used for data understanding, feature screening, and preprocessing, stages that strongly influence downstream results. In the preprocessing phase, in particular, a key step is to remove outliers, i.e., data that significantly deviates from other instances in the dataset. If these outliers are included during the model training phase, they can negatively impact the model's performance or introduce bias. A commonly used method for outlier detection is the Isolation Forest (iForest), a tree-based ensemble model that is computationally easy to train [143].

However, tree-based ensemble models are not transparent by default when scaled up. Large ensembles with hundreds of trees are hard to interpret, and their decision process remains opaque [125]. Common summaries based on Feature Importance (FI) can also mislead, especially when features do not have the same scale of measurement or when they are correlated [256]. The central challenge is to extract explanations that are informative and operationally meaningful from models that are both accurate and practical under industrial conditions.

The food sector increasingly needs AI applications that accelerate and simplify quality analysis while improving product quality. These applications must be reliable and transparent not only at the point of prediction but across the entire workflow, from dataset analysis to preprocessing through to the final output. What the industry needs is an end-to-end explainable AI pipeline.

1.2 Contributions

In this thesis, we propose a set of techniques to build a fully explainable ML pipeline based on tree-based ensemble models for tabular data in food engineering.

Working with experts at ASAC srl, we examined the food supply chain. We identified food quality analysis as the step with both the strongest need and the best opportunity to introduce transparent AI that improves process performance. The project began with a review of the literature to assess the state of the art in explainable AI for food quality analysis and to recognize open issues. This review work proved extremely complex due to the low number of published works and, above all, the lack of a comprehensible taxonomy to serve as a key for reading and organizing scientific articles. Furthermore, since these articles concerned the applications of algorithms and technologies specific to the field of computer science to the world of food engineering, the articles mentioned above were not homogeneous in the presentation of the data, in the discussions of the results, and, in general, in the nature of the academic contribution. Our first contribution is a structured literature review that organizes existing work on explainable AI for food quality analysis. Within this review, we introduce a taxonomy that classifies studies by the elements and challenges specific to food quality analysis, for example, safety parameters or nutritional values assessment. This catalog provides both food quality specialists and computer scientists with a shared reference point, aiming to narrow the gap between the fields and encourage further development and deployment of explainable AI in food quality analysis.

From the literature review, we found that XAI is widely applied to pictorial data, while its use with tabular data is still in its early stages. Despite this, tabular data is both valuable and prevalent in food engineering because it is easy to acquire, often automatically, and has the ability to convey diverse features. Furthermore, tabular data allows for the use of tree-based ensembles, which, as we previously mentioned, come with significant advantages.

While numerous explainability methods exist for tree-based ensemble models, most provide *local* explanations, i.e., explanations that clarify the model's decision-making process for individual samples [69]. While local methods are useful for understanding how the model arrives at decisions for specific cases, they do not offer a comprehensive view of the overall decision-making process. That role is fulfilled by *global* methods, which aim to provide insights into the model as a whole. In contexts like food quality analysis, where it is important to establish stable criteria at the model level for evaluating effectiveness across different batches and product lines, global explanations can be more beneficial. Additionally, understanding the model's overall decision-making is relevant for identifying errors, recognizing biases, and confirming the reliability of results, regardless of specific cases.

Recent proposals for techniques that provide global explanations, specific to tree-based ensemble models, focus on simplifying decision structures through graph visualization and emphasizing key paths. However, while these efforts enhance the visualization experience, they may either result in a visually complex representation as the number of tree-based learners increases [89, 182] or compromise the interpretability of the original ensemble model due to the need for pruning or feature

selection [270, 310].

Consequently, the second and most important contribution of this thesis is the development of a new explainability method, called Decision Predicate Graph (DPG), specifically designed for tree-based ensembles, which provides a global explanation. DPG represents the ensemble and learned dataset details as a graph that preserves relationships among features, logical decisions, and predictions, highlighting the most informative elements. Leveraging well-known graph theory concepts, such as centrality measures and community structure, DPG offers additional quantitative insights into the model, complementing visualisation techniques, expanding the problem space descriptions, and offering diverse possibilities for extensions. Empirical experiments demonstrate the potential of DPG in addressing traditional benchmarks and complex classification scenarios.

Although DPG has demonstrated its effectiveness in explaining various tree-based ensemble models during its development, for this doctoral thesis, we study its application to one of the most widely used and significant models, namely RF. Therefore, in the following chapters, we will explore this model in detail.

After developing DPG, we explored its use in a significant and sensitive area of food quality analysis: ensuring transparency and trust in the fruit supply chain is necessary to improve food safety, reduce waste, and maintain high-quality standards. To meet this need, we combined physicochemical descriptors with tree-based ensemble models to predict fruit quality and detect potential problems such as spoilage. We introduced an approach that uses explainable AI to clarify models. Specifically, we proposed an application to the analysis of different ripening stages of three distinct fruits: pitaya (dragon fruit), carambola (star fruit), and papaya, using tabular physicochemical data. Our goal was to deliver both global and local explanations for ensemble methods by pairing DPG with Building Explanations through a LocalLy AccuraTe Rule EXtractor (BELLATREX), a well-established technique in the literature. Together, these methods form a suite of ensemble-based explainability tools for the fruit supply chain. By integrating these techniques, we improve transparency and reliability while deepening understanding of relationships among food quality parameters. The approach highlights key attributes that influence fruit quality and helps detect potential fraud in the supply chain, addressing traceability challenges.

Beyond making the model interpretable, careful data preprocessing is equally necessary. Understanding how transformations affect performance and bias not only improves results but also builds a robust and reliable pipeline. This is where truly resilient AI solutions take shape. For this reason, a central part of this doctoral work was to explain a key preprocessing algorithm as part of a fully explainable AI-based pipeline.

iForest, which we discussed earlier, is a tree-based ensemble model widely used for identifying outliers. As with other ensembles, its effectiveness often increases with the number of learners. However, this enhancement can make it more challenging to understand how outliers are determined. Existing explainability methods for these models tend to provide local views and do not clarify the overall decision-making process.

In this context, while it is sometimes preferable to understand the logic underlying the identification of a single case, the ensemble structure of iForest suggested

an opportunity to design a dedicated method for global insight. We therefore introduced a new explainability approach that extends the DPG to iForest, addressing the problem of global explainability. It offers a comprehensive view of how the model detects outliers by indicating which features drive the decision and how they are used. The method advances the state of the art by illuminating decision boundaries and revealing the holistic role of features in outlier detection.

The final step of this doctoral project is to demonstrate the effectiveness of the proposed tools by building a fully explainable end-to-end pipeline based on AI. For this step, we examined an application outside of food engineering due to the time constraints involved in generating a dataset that met all experimental requirements—specifically, one that included relevant outliers, reflected a well-studied problem, and remained computationally manageable. While food-related datasets are conceptually easy to obtain, their construction requires sensing equipment and acquisition time. Therefore, to ensure a reliable and interpretable validation of the entire pipeline within the project's time, another benchmark was selected.

We developed an end-to-end pipeline for finance, a high-risk domain that requires finding patterns and insights in complex and heterogeneous data. The scenario is similar to food quality analysis, as the needs are the same in both situations: the opacity of many AI models raises concerns about interpretability and reliability, especially in financial decision-making. Furthermore, real-world financial data is often vulnerable to outliers due to data entry errors, fraud, or rare but critical events, which can harm model performance and regulatory compliance.

We therefore proposed a framework that explains the entire ML pipeline in a financial risk assessment scenario. Specifically, we applied DPG methods to two key steps in the pipeline: outlier cleansing with iForest during preprocessing, and the classification for credit scoring with RF. A real-world case study shows that the approach provides clear insight into both preprocessing and predictive components, improving reliability and strengthening user confidence. The proposed example can be extended to more complex applications in food engineering, enabling an end-to-end explainable pipeline for the food supply chain and supporting food quality analysis.

To summarize, the main contributions of this thesis are:

- A structured literature review and taxonomy of XAI for food quality analysis, which bridges the fields of food engineering and computer science, and organizes sparse, heterogeneous studies into a usable reference for both domains.

- A novel global explainability method for tree-based ensembles, called DPG, which provides quantitative and visual insights into the model's decision-making process.

- A case study in the fruit supply chain using physicochemical tabular data (including pitaya, carambola, and papaya) that combines DPG (global) with BELLATREX (local), offering a complementary suite of "glocal" explanations for quality assessment.

- An extension of DPG to providing a global explanation of iForest, clarifying which features influence outliers detection.
- Guidelines for explainable preprocessing that make data transformations interpretable, demonstrating how transparency during the cleaning and detection stages enhances the overall performances of the pipeline.
- An end-to-end explainable ML pipeline demonstrated in a financial risk assessment setting (iForest for cleansing and RF for credit scoring with DPG explanations), evidencing improvements in reliability and user trust in a high-risk domain.
- Actionable pathways for the deployment of transparent, ensemble-based XAI in the food supply chain, informed by collaboration with domain experts (ASAC srl) and designed to support real-world quality, safety, and traceability objectives.

1.3 Structure of the Thesis

To facilitate a clear understanding of the thesis and demonstrate how we achieved the research objectives outlined in Chapter 1, we have organized the content as follows.

- Chapter 2 provides a comprehensive explanation of the fundamental concepts and tools utilized throughout this thesis. The Chapter begins by defining the principles of ML, the learning paradigms relevant to this project, and the evaluation metrics used to assess the performance of the developed ML techniques. The Chapter then explores ensemble methods, focusing on Decision Trees, RF, and iForest. Finally, we present a taxonomy for identifying and classifying XAI techniques, including a distinction between different types of explanations, followed by an overview of the most commonly used models.
- Chapter 3 first introduces the taxonomy of food quality analysis tasks that we created to organize scientific publications related to XAI applications in the field of food quality analysis. It also details how the papers are arranged. This organization includes, in addition to the taxonomy, the type of data used and the type of explanations generated by the XAI techniques adopted in each paper. The Chapter contains over one hundred collected papers and constitutes an overview of the state of the art. Finally, the Chapter highlights the gap between food quality analysis and explainability.
- Chapter 4 introduces DPG, a new global and model-specific explanation for tree-based ensembles. The Chapter formalizes the concept of DPG, presents the algorithm for its construction along with its complexity, and demonstrates how graph-theoretic tools can reveal the ensemble's global decision-making logic. Additionally, it provides empirical demonstrations, compares DPG to other graph-based alternatives, and discusses its limitations and potential improvements.

- Chapter 5 focuses on the food domain, specifically examining XAI in the fruit supply chain. The Chapter proposes a study that assesses the level of fruit ripeness using tabular physicochemical characteristics to classify the fruit and determine its quality. It further describes the proposed methodology and the datasets used, including star fruit, papaya, and pitaya. The Chapter pays particular attention to the two XAI methods employed: DPG and BELLATREX. These methods provide different types of explanations—global and local, respectively. When used together, they provide a comprehensive understanding of the decision-making process of the AI method, thereby ensuring its transparency.

- Chapter 6 highlights the importance of explaining the preliminary stages of an AI application, particularly the preprocessing phase. It extends DPG to the iForest method, which is aimed at providing global explainability for outlier detection. The Chapter introduces a DPG-based explanation specifically designed for iForest and presents the Inlier-Outlier Propagation Score (IOP-Score) to assess how various predicates contribute to the isolation process. Finally, the method is evaluated using both synthetic and benchmark data.

- Chapter 7 composes all the developed and studied techniques into an end-to-end explainable pipeline in a high-stakes finance setting. This pipeline is validated through its application in credit-risk assessment, detailing the preprocessing steps, which include iForest-based outlier cleaning using DPG-extended technique, and the predictive modeling phase, which employs a RF algorithm with DPG. Additionally, the Chapter outlines the choices made in dataset selection and provides experimental details along with the results obtained.

- Chapter 8 concludes with a summary of contributions, limitations, and future directions. It consolidates how the taxonomy, DPG methods, and pipeline enhance reliable, global explainability for tree-based ensembles and outlines future research, including applications to upcoming food datasets.

In Figure 1.1, the timeline structure of the thesis is highlighted, with its central Chapters shown.

1.4 List of Publications

During my PhD, I produced the following publications that serve as the foundation for this thesis.

- [10] L. Arrighi, L. Pennella, G. Marques Tavares, and S. Barbon Junior. Decision predicate graphs: Enhancing interpretability in tree ensembles. In *World Conference on Explainable Artificial Intelligence*, pages 311–332. Springer Nature Switzerland, 2024. ISBN 978-3-031-63797-1. doi: 10.1007/978-3-031-63797-1_16

- [39] M. Ceschin, L. Arrighi, L. Longo, and S. Barbon Junior. Extending Decision Predicate Graphs for Comprehensive Explanation of Isolation Forest. In

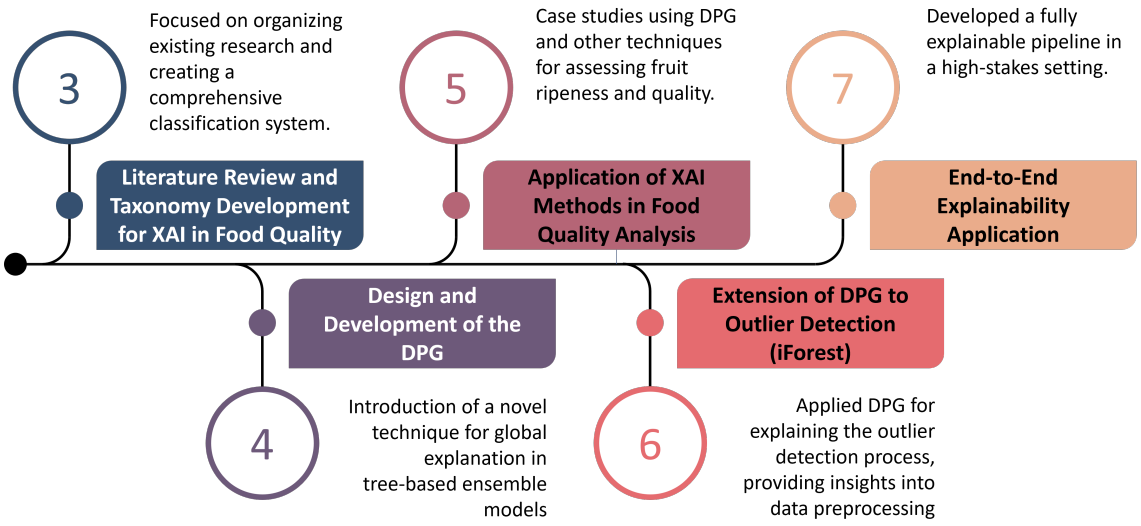


Figure 1.1: Representation of the thesis structure, highlighting the central Chapters enclosed in circles, with a brief summary of each Chapter provided next to the corresponding numbers.

R. Guidotti, U. Schmid, and L. Longo, editors, *Explainable Artificial Intelligence*, pages 271–293. Springer Nature Switzerland, 2026. ISBN 978-3-032-08324-1. doi: 10.1007/978-3-032-08324-1_12

The paper "Extending Decision Predicate Graphs for Comprehensive Explanation of Isolation Forest" won the Best Student Paper award at World Conference on eXplainable Artificial Intelligence 2025 in Istanbul, Turkey.

- [11] L. Arrighi, M. Camilo Da Silva, and S. Barbon Junior. End-to-End Explainability of Machine Learning Pipelines with Decision Predicate Graphs: A Financial Scenario Case Study. In *Ital-IA 2025 CINI National Conference on Artificial Intelligence*, 2025

- [12] L. Arrighi, I. A. de Moraes, M. Zullich, M. Simonato, D. F. Barbin, and S. Barbon Junior. Explainable artificial intelligence techniques for interpretation of food datasets: A review, 2025. URL https://arxiv.org/abs/2504.10527. *(Under Review)*

- [13] L. Arrighi, C. Giaccari, I. A. de Moraes, D. F. Barbin, and S. Barbon Junior. Enhancing Transparency in the Fruit Supply Chain Using eXplainable Artificial Intelligence, 2025. *(Under Review)*

I also created the following publications that are not included in this thesis.

- [9] L. Arrighi, S. Barbon Junior, F. A. Pellegrino, M. Simonato, and M. Zullich. Explainable Automated Anomaly Recognition in Failure Analysis: is Deep Learning Doing it Correctly? In *Explainable Artificial Intelligence*, Communications in Computer and Information Science, pages 420–432. Springer Nature

Switzerland, 2023. ISBN 978-3-031-44067-0. doi: 10.1007/978-3-031-44067-0_22

- [61] I. A. de Moraes, L. Arrighi, S. Barbon Junior, J. E. L. Villa, R. L. Cunha, and D. F. Barbin. Explainable artificial intelligence (xAI) applied to deep computer vision of microscopy imaging and spectroscopy for assessment of oleogel stability over storage. *Journal of Food Engineering*, 394:112515, 2025. ISSN 0260-8774. doi: 10.1016/j.jfoodeng.2025.112515

- [178] I. A. Moraes, L. Arrighi, S. B. Junior, R. L. Cunha, and D. F. Barbin. Explainable artificial intelligence (XAI) applied to deep computer vision for the assessment and classification of oleogels with varying oleogelator types and concentrations. *Microchemical Journal*, page 116821, 2026. ISSN 0026-265X. doi: 10.1016/j.microc.2026.116821

- [14] L. Arrighi, I. A. de Moraes, M. Simonato, and S. Barbon Junior. Discriminating Short-Term Moisture Changes in Stuffed Pasta Using Deep Computer Vision. In E. Rodolà, F. Galasso, and I. Masi, editors, *Image Analysis and Processing - ICIAP 2025 Workshops*, pages 489–496, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-11381-8. doi: 10.1007/978-3-032-11381-8_40

1.5 Data and Code Availability Statement

All software libraries and datasets used in this thesis are publicly available or can be provided upon request. Specific links to the libraries are given in footnotes at the point where they are mentioned. A consolidated list of the resources and analysis code, including repository links or identifiers, is provided below.

- DPG: https://github.com/LeonardoArrighi/DPG
- Synthetic dataset used for testing DPG: https://github.com/LeonardoArrighi/DPG/tree/main/datasets
- DPG-extension for iForest: https://github.com/Math0097/DPG-iforest
- Synthetic dataset used for testing the end-to-end explainable framework: https://github.com/LeonardoArrighi/DPG-Pipeline

Chapter 2

Conceptual Framework

In this Chapter, we introduce the definitions, concepts, and algorithms that ground the thesis. Section 2.1 presents core ML theory and the evaluation metrics used to assess the models developed in the subsequent Chapters. Section 2.2 introduces the tree-based algorithms employed extensively. Finally, Section 2.3 outlines the key concepts of XAI, along with the classification taxonomy used to describe these techniques, and provides a general overview of some widely used XAI methods. As general background, the exposition is informed by [26, 176, 181, 240], which serve as principal references.

2.1 Base of Machine Learning

ML is a branch of AI that discovers patterns or regularities in data and uses them to make predictions or decisions. Instead of being explicitly programmed for each task, an ML system learns a mathematical model from experience, transforming data into a parametric model that makes predictions on new, unseen cases. A good model can generalize, i.e., it inductively transfers the regularities discovered in the observed data to new data drawn from the same (or a related) distribution.

Let the dataset be $\mathcal{D} = \{z_i\}_{i=1}^{m}$. A model is a parameterized function f_ω with free parameters ω. *Training* is the process of selecting ω by minimizing an objective (loss) function on $\mathcal{D}$, to promote generalization to unseen data. During training, *overfitting* may occur when the model captures noise or incidental structure in $\mathcal{D}$, achieving low training error but performing poorly on new samples. On the contrary, *underfitting* occurs when the model is too limited to represent the underlying signal.

To perform training, a data-splitting procedure is applied to both optimize the model and assess its performance. The dataset is partitioned into three subsets: a training set, whose data are used to fit the model; a validation set, whose data are used to evaluate the fitted model by comparing its predictions to the expected outputs and to guide model selection and hyperparameter tuning; and a test set, whose data are used to estimate the model's generalization performance. The training and validation sets are used during model development, whereas the test set remains hidden throughout training and model selection so that the final evaluation provides an unbiased assessment.

A central consideration is the trade-off between model complexity and available

data. Model capacity, determined by the richness of the hypothesis class and the number of parameters, must be commensurate with the size and quality of the dataset. Excess capacity relative to data increases variance and the risk of overfitting, while insufficient capacity yields bias and underfitting. This balance is managed by aligning architecture and parameterization with dataset scale and by employing regularization and rigorous validation.

2.1.1 Supervised Learning and Unsupervised Learning

Supervised and *unsupervised* are two popular learning paradigms in ML.

In supervised learning the data are labeled, $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^m$, with inputs $x_i \in \mathcal{X}$ and targets $y_i \in \mathcal{Y}$. The inputs may be represented as feature vectors or as objects, like pictures. At the same time, the targets define the nature of the task, being real-valued for *regression* and categorical for *classification*. Training minimizes a loss that measures the discrepancy between predictions and ground-truth labels. In this thesis, various supervised learning algorithms are discussed. This includes tree-based methods, such as RFs [31] and gradient-boosted trees like eXtreme Gradient Boosting algorithm (XGBoost) [47], along with k-Nearest Neighbors algorithm (kNN) and support vector machines extensively described in [101]. Additionally, various Neural Network (NN) architectures are covered, with a particular focus on Convolutional Neural Network (CNN), including prominent architectures such as Residual Network [103], Visual Geometry Group (VGG) [251], and You Only Look Once (YOLO) [221].

In unsupervised learning the data are unlabeled, $\mathcal{D} = \{x_i\}_{i=1}^m$ with $x_i \in \mathcal{X}$. The goal is to learn the underlying structure or a model of the data distribution by optimizing an objective defined solely on the inputs, without access to target labels. Training and evaluation, therefore, operate on the same input format, and success is measured by how well the learned representation or model captures salient regularities in $\mathcal{X}$. This thesis also mentions unsupervised learning algorithms, including clustering methods described in [101] and various DL techniques such as Generative Adversarial Network (GAN) [88], autoencoders [228], and visual transformers [65].

Other learning paradigms lie outside the scope of this manuscript.

2.1.2 Metrics

To evaluate the performance of ML models and enable fair comparisons on a common dataset, we report a set of standard metrics.

In multiclass setting (with classes $\{1, \ldots, C\}$), let the *confusion matrix* be $\mathbf{C} \in \mathbb{N}^{C \times C}$, where rows index true classes and columns index predicted classes, so that C_{ij} counts instances with true class i predicted as j.
For a given class c, define *True Positive* (TP), *Negative* (N), *False Positive* (FP), *False Negative* (FN):

$$\mathrm{TP}_c = C_{cc}, \quad \mathrm{FP}_c = \sum_{i \neq c} C_{ic}, \quad \mathrm{FN}_c = \sum_{j \neq c} C_{cj}, \quad \mathrm{N}_c = \sum_{i=1}^{C} \sum_{j=1}^{C} C_{ij}.$$

Hence, we can define the metrics:

Accuracy. The proportion of correctly classified instances:

$$\text{Accuracy} = \frac{\sum_{c=1}^{C} C_{cc}}{N}.$$

Precision. The fraction of predictions for class c that are correct:

$$\text{Precision}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c} = \frac{C_{cc}}{\sum_{i=1}^{C} C_{ic}}.$$

Recall. The fraction of true instances of class c that are correctly identified:

$$\text{Recall}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c} = \frac{C_{cc}}{\sum_{j=1}^{C} C_{cj}}.$$

F1-Score. The harmonic mean of precision and recall for class c:

$$\text{F1}_c = \frac{2\,\text{Precision}_c\,\text{Recall}_c}{\text{Precision}_c + \text{Recall}_c}.$$

Average F1-Score. Per-class F1-scores can be aggregated as:

$$\text{F1-Score} = \frac{1}{C}\sum_{c=1}^{C} \text{F1}_c.$$

2.2 Ensemble Methods

Ensemble models address a common problem in predictive modelling. A single model can be accurate on average yet fragile on new data. By combining many simple learners that make different mistakes, an ensemble lets complementary strengths accumulate while individual errors cancel out. The result is a predictor that is both accurate and dependable, with reduced variance and stronger generalization.

In the following sections, we outline the tree-based building blocks used throughout this thesis, namely binary trees and decision trees. In the following sections, we outline the tree-based building blocks used throughout this thesis, namely binary trees and decision trees. Additionally, we will introduce the two ensembles of interest, which are the main models studied in this thesis and are employed to demonstrate the techniques we have created and developed. A RF averages the outputs of many randomized trees and stabilizes predictions. An iForest identifies outliers by measuring how quickly observations become isolated within a collection of binary trees.

2.2.1 Binary Tree and Decision Tree

A *binary tree* is a finite collection of nodes connected by directed edges, with each node having at most two children. The first node of the structure is called the *root*, and any node directly below another is a *child*. A binary tree is defined recursively, either the structure is empty or it consists of a distinguished root together with a left and a right subtree, each of which is itself a binary tree. Nodes with no children

are called *leaves*, and nodes with at least one child are *internal*. The *depth* of a node is its distance from the root, and the *height* of a tree is the maximum depth among its leaves. This recursive form supports divide-and-conquer procedures and traversal algorithms. In statistical learning, binary trees are beneficial for partitioning a space. Each internal node applies a rule that splits the current set into two parts, and the leaves summarize the outcomes on the resulting subsets.

A *decision tree* is a supervised learning model structured as a binary tree. Each internal node evaluates a feature value test, and each leaf encodes a prediction. In classification, leaves return either a class label or a vector of class probabilities; such models are called *classification trees*. In regression, leaves return a real-valued prediction.

Training proceeds recursively. At each node, the algorithm searches over candidate splits, namely thresholds for numerical features and subset splits for categorical features, and partitions the current sample into two child nodes. The procedure is repeated on each child until a stopping condition is met. Stopping criteria include a maximum tree depth and a minimum sample size required to attempt a split.

Split selection is guided by an impurity criterion that quantifies node heterogeneity, interpreted as the probability that a randomly chosen instance would be incorrectly labeled. Let y denote the labels at a node and C the set of classes. If $p_{y,c}$ is the proportion of class c in that node, the Gini impurity (G) is

$$G(y) = \sum_{c \in C} p_{y,c} \big(1 - p_{y,c}\big),$$

which ranges from 0 for a pure node to $(|C|-1)/|C|$ for a uniform class distribution. A convenient way to compute the proportions is

$$n_c = \sum_{i=1}^{|y|} \mathbf{1}\big(y_i = c\big), \qquad p_{y,c} = \frac{n_c}{|y|},$$

where $\mathbf{1}(\cdot)$ is the indicator function. For a candidate split with left and right children L and R having sizes n_L and n_R from a parent of size n, the quality of the split is measured by the impurity reduction

$$\Delta = G(y_{\text{parent}}) - \frac{n_L}{n} G(y_L) - \frac{n_R}{n} G(y_R),$$

and the algorithm chooses the split that maximizes Δ.

The overall objective is to choose, at each step, the split that produces the most significant reduction in impurity.

2.2.2 Random Forest

A RF is a tree-based ensemble learning method that extends the decision tree model by combining an extensive collection of trees. The central idea, introduced by [31], is that averaging many weakly correlated predictors reduces variance and leads to a more accurate and stable model.

The construction of a RF proceeds in two stages. Given a fixed number of trees, each tree in the forest is grown on a bootstrap sample, i.e., a sample of the training observations drawn with replacement and of the same size as the original dataset. This introduces variability among the trees. In addition, at each node, rather than considering all available features for candidate splits, only a random subset of features is evaluated. This second source of randomness reduces correlation among trees. Each tree in the forest is trained independently on its own bootstrapped dataset. RFs are grown to maximum depth, so that individual trees exhibit low bias and high variance. Averaging across trees reduces variance, and the randomization of feature selection reduces correlation, thereby improving generalization.

Formally, let $T_b(x)$ denote the prediction of the b-th tree, grown from bootstrap sample Z_b and with feature subsampling at each split. For regression, the RF predictor is the ensemble average

$$\hat{f}(x) = \frac{1}{B} \sum_{b=1}^{B} T_b(x),$$

where B is the total number of trees. For classification, the forest prediction is obtained by majority vote among trees,

$$\hat{C}(x) = \arg\max_{c \in C} \sum_{b=1}^{B} \mathbf{1}(T_b(x) = c),$$

where C is the set of classes and $\mathbf{1}(\cdot)$ is the indicator function. Equivalently, one can average the estimated class probabilities across trees and choose the class with the highest average probability.

RFs provide two valuable by-products in addition to accurate predictions. First, they yield an internal estimate of generalization error, known as the Out-Of-Bag (OOB) error. Because each tree is built on a bootstrap sample, roughly one-third of the training observations are not included in that sample. The tree can predict these excluded observations, and averaging over trees gives an unbiased estimate of test error without the need for a separate validation set. Second, RFs provide measures of variable importance. The most common approach is to evaluate the increase in prediction error when the values of a given feature are permuted across the OOB samples, thereby breaking the association between that feature and the response.

From a theoretical perspective, RFs reduce variance by averaging across many deep decision trees while controlling correlation through feature randomization. As argued by [101], the resulting model is highly flexible, able to capture complex nonlinear relationships and interactions, while remaining robust to noise and overfitting.

2.2.3 Isolation Forest

One of the most widely used algorithms for outliers detection is iForest, a tree-based method introduced by [143]. iForest is designed to efficiently identify outliers—data points that deviate significantly from other instances in the dataset—rather than inliers, which represent most of the data and conform to expected patterns. Among the various techniques available, iForest stands out for its efficiency and scalability,

thanks to its linear time complexity and low memory consumption. Another key advantage is that iForest is an unsupervised learning method that does not require labelled data for training. Moreover, through an effective subsampling procedure, iForest mitigates the swamp effect, where regular points are wrongly identified as outliers, and addresses the masking issue, which occurs when multiple outliers conceal each other. iForest identifies outliers by recursively partitioning the data. Its core idea is that these irregular data points are rare and distinct from normal instances, requiring fewer random splits to isolate in the problem space. This characteristic enables the algorithm to efficiently separate outliers from the majority of inliers.

Given a dataset X, where d features characterize each instance, the iForest consists of multiple binary trees, called Isolation Tree (iTree), that form the forest. Each tree is built by randomly selecting a feature d_i and a random value v within the range $[\min(v_{d_i}), \max(v_{d_i})]$, where v_{d_i} are the values of the samples of X associated to the feature d_i. If an instance's selected feature value v_{d_i} is less than v, the instance is directed to the left branch; otherwise, it is directed to the right branch of the iTree. After each split, the dataset is partitioned so each branch contains a subset of X. This process is recursively applied to the resulting subsets until one of the following stopping conditions is met:

- The iTree reaches its maximum depth, which is defined as:

$$\lceil \log_2(\min(256, |X|)) \rceil,$$

  where $|X|$ is the number of samples of the dataset. This ensures that the tree does not grow indefinitely.

- A single instance has been completely isolated in a leaf node.

- Two or more identical instances have been grouped into a single leaf node, making further splits impossible.

Once an iTree is fully grown, each instance x in X is assigned to a leaf node. Its path length $h(x)$ is the number of edges traversed from the root to that leaf. This recursive process is repeated n times to build n trees in the forest. The final step of the iForest algorithm is the calculation of the *anomaly score* for each instance in the dataset. This score allows the model to determine whether a sample is an outlier (anomaly) or an inlier. The anomaly score is computed as follows:

$$s(x, n) = 2^{-\frac{\mathbb{E}(h(x))}{c(n)}},$$

where $\mathbb{E}(h(x))$ is the average path length of x across all trees in the forest, and $c(n)$ is a normalization factor that estimates the average path length required to isolate a data point in a binary search tree containing n instances and is given by:

$$c(n) = 2H(n-1) - \frac{2(n-1)}{n},$$

where $H(i)$ is the harmonic number, and it can be estimated by $ln(i) + 0.5772$ (Euler's constant). If $s(x, n) < 0.5$, then x is likely to be a typical instance (inlier). Conversely, if $s(x, n)$ is close to 1, then x is highly likely to be an outlier. The core idea behind the iForest algorithm is that outliers require fewer partitions to be isolated, resulting in shorter path lengths than inliers.

2.3 Explainable Artificial Intelligence

Explainability and *interpretability* in the context of AI models, although often used interchangeably, have slightly different meanings, as elucidated by [32]. The authors argue that interpretability is concerned with understanding the inner workings of a model. In contrast, explainability is strictly tied to providing *post-hoc*, approximate insights on a *prediction* operated by the model. To clarify, the term *post-hoc* refers to the analysis conducted after the model has been trained.

In other words, interpretability is an intrinsic property of a model, while an explanation is generated on a (non-)interpretable model after a prediction has been made.

Accuracy vs. interpretability trade-off: Interpretability, as defined above, has often been depicted as being at odds with *accuracy*[1], also termed *expressivity* or *flexibility* which we prefer to call, of the model [115]. Flexibility refers to the range of complicated patterns that the model can learn.

Linear Regression (LR) is often depicted as a very inflexible model because it can only learn simple linear relationships between predictors; hence, its accuracy will be fairly limited on more complex problems, such as those involving pictorial data. However, the linear relationship is interpretable by human standards: a single parameter of an LR model indicates the additive effect that a perturbation of the corresponding predictor has on the response. This makes it straightforward to analyze, for instance, the importance of each variable within the model.

Conversely, highly expressive models such as Deep NNs are considered complex. While they achieve high accuracy on very intricate problems, interpreting the rules these models use to make specific predictions can be quite challenging. A depiction of this trade-off can be seen in Figure 2.1, where some of the most popular AI models are positioned accordingly. Despite the trade-off being renowned in the literature, it is still an approximate rule-of-thumb, which has exceptions, like in the case of Vision Transformer (ViT)s [66], which, despite being more flexible than CNNs, are defined as inherently more interpretable due to the ease of visualizing the attention mechanism [244].

Global vs. local XAI methods: Another axis that defines XAI tools is represented by the *scope* of the method. If the tool delves into properties of the model as a whole, then the scope is said to be *global*; conversely, when the tool investigates the model behavior around one data point, then the scope is said to be *local*. Recently, [145] introduced the term *glocal* to describe a method that combines a local XAI approach with a global one. Concerning the LR example before, the model's coefficients can be thought of as global explanations, since they define a global behaviour

[1] In this specific case, we use the term *accuracy* as a generic stand-in for the performance of the model in solving the task which it was designed to carry out.

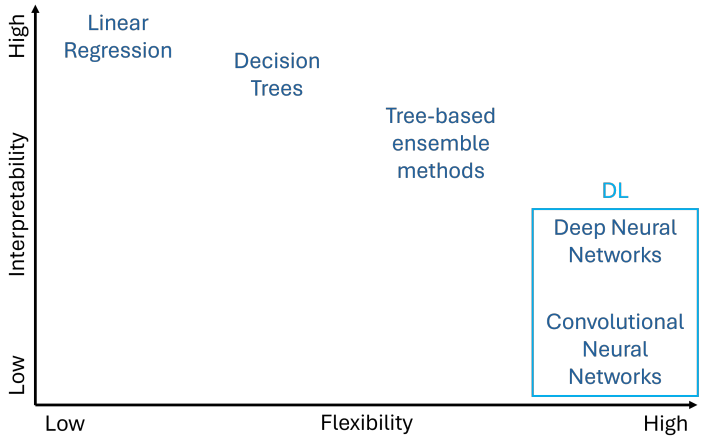


Figure 2.1: Chart illustrating approximately the trade-off between expressivity or flexibility and interpretability. Expressive models, such as those based on DL, are capable of reaching higher task-level performance but are often difficult to interpret. On the other hand, less complex models, like LR, are inherently interpretable, but often incapable of attaining high task-level performance.

of the model irrespective of the specific data point considered. On the other hand, as an example of local explanation, we can consider *feature attribution* in the context of image classification using CNNs. For feature attribution, we indicate the action of identifying which variables contribute the most to producing the prediction. In the case of image classification, it may be of interest to elicit *important* pixels that led a given picture to be classified in a given category; this is an example of a local explanation since we are gaining knowledge of the behavior of the model only on the current image, without trying to infer the global properties of the model. In the case of NNs, it is often hard to identify such global rules for explaining predictions; thus, local explanations are often preferred [287]. Despite being limited in scope, local explanations can be used to extrapolate global information about the models, as, for instance, in the works by [287] and [236].

Model-agnostic vs. model-specific XAI methods: A final property of the XAI tools to be considered is the *specificity* to limited classes of models. *Model-agnostic* tools are XAI methods that, due to how they are constructed, can be applied to any AI model, while *model-specific* tools are restricted to limited classes of models.

2.3.1 Explanation Types

A particularly informative classification considers XAI techniques in terms of how explanations are provided to the user. Because much of XAI's value resides in the explanation format, this perspective helps distinguish models and clarify what they actually do. Vilone et al. [274] exemplify this approach by classifying methods by output modality—*numerical*, *rule-based*, *textual*, *visual*, or *mixed*—as shown in Figure 2.2. Selecting among these modalities depends on the use case, as distinct scenarios require different methods to clarify model behavior.

Numerical explanations: Numerical explanations are defined as the conveying

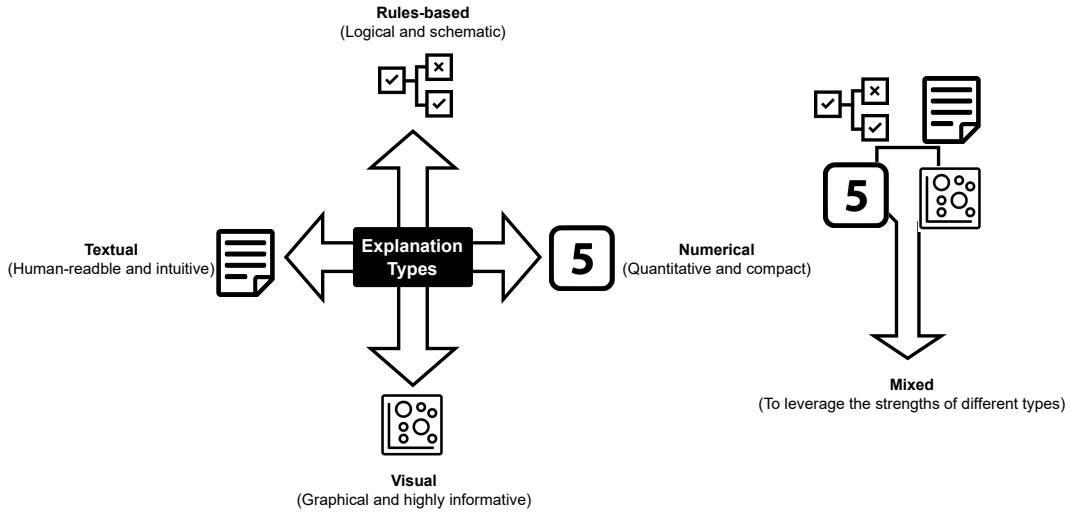


Figure 2.2: Representation of the types of explanations provided by XAI techniques, along with a summary of their key advantages.

of information in a compact format using crisp values, vectors of numbers, matrices, or tensors to highlight the input attributes or features of a model that have the most significant effect on the prediction of the output.

Visual explanations: Visual explanations use graphical tools to illustrate information, often through heatmaps, graphs, or other visualizations that highlight specific areas of the data that influence the model's inferential process.

Rule-based explanations: Rule-based explanations use a logical format, typically in the form of "IF... THEN" statements with AND/OR operators, to express combinations of input features and their activation values. These rules employ symbolic logic, a formalized system of primitive symbols and their combinations.

Textual explanations: Textual explanations consist of natural language statements that can either be written or orally uttered, providing a human-readable and intuitive format for explanations.

Mixed explanations: Mixed explanations combine multiple formats, such as visual, textual, and numerical explanations, to exploit their strengths and overcome individual weaknesses.

2.3.2 Overview of Popular Explainable Artificial Intelligence Techniques

In recent years, several explainability techniques have gained importance in response to the growing need for transparent AI models. These methods are generally straightforward to apply and support a range of models and explanation types. Together, they constitute a versatile toolkit that can be adapted to diverse cases and operational contexts. In this Section, we do not detail every technique because of space constraints. Instead, we present a concise summary that highlights the key aspects, indicates the explanation type for each method, and provides a general classification based on the taxonomy introduced in the preceding section. An overview

Table 2.1: Overview of the most popular XAI methods classified by scope, specificity, and explanation type.

XAI technique	Scope	Specificity	Explanation type
LIME [223]	Local	Model-agnostic	Numerical / Mixed
SHAP [148]	Local	Model-agnostic	Numerical / Mixed
PDP [81]	Global	Model-agnostic	Visual
LRP [18]	Local	Model-specific	Visual
CAM [308]	Local	Model-specific	Visual
Grad-CAM [234]	Local	Model-specific	Visual

of the methods is reported in Table 2.1.

LIME [223] explains a single prediction by locally imitating the black box of a model. It involves perturbing the instance, querying the model on these nearby samples, and weighting them based on their proximity. A simple and interpretable model, such as a sparse linear model or a small decision tree, is then fitted to this weighted neighborhood. The coefficients of this model reveal the local factors that influence the decision. While the explanation generated by LIME is numerical, it is often represented visually as a bar plot.

SHAP [148] explains a single prediction by attributing to each feature its game-theoretic contribution to the model's output. It compares the prediction for the instance with predictions where features are treated as missing using a background dataset, then aggregates each feature's marginal effect across many coalitions to compute its SHAP value. The explanation is additive, so the sum of feature contributions equals the difference between the model's expected output and the instance's prediction. While the explanation is numerical, it is typically visualized with dedicated attribution plots.

PDP [81] explains model behavior by showing the marginal effect of one or two features on the predicted outcome. They are computed by varying the selected feature values over a grid, querying the model at each value while averaging predictions over the distribution of all other features, and assembling the resulting partial dependence function. This isolates the average relationship between the chosen features and the prediction, revealing trends such as monotonicity, saturation, or interaction strength. The explanation consists of a line plot for one feature or a surface plot for two features.

LRP [18], CAM [308], and Grad-CAM [234] explain a single prediction by tracing class evidence back to the input. LRP backpropagates the output score through the network with relevance conservation, redistributing relevance at each layer in proportion to local contributions to yield a pixel or feature relevance map. CAM and Grad-CAM compute a class-specific localization map by weighting the final convolutional feature maps with the learned class weights and aggregating them to highlight discriminative regions. Both methods generate numerical attributions that are typically represented as heatmaps over the input. They are model-specific. Although they can explain a broad range of NN architectures, they are designed primarily for CNNs.

Chapter 3

Explainable Artificial Intelligence in Food Quality Analysis

In this Chapter, we present an analysis of XAI applications in food engineering, with the goal of mapping how explainability enhances transparency in AI-driven food quality assessment. This analysis considers works published up to December 2024. Section 3.2 introduces the proposed taxonomy, describes the criteria used to classify the studies, and summarizes the AI methods employed in the corpus. In Sections 3.3 to 3.7, we analyze the selected works, detailing the XAI techniques adopted, the associated AI tasks, and the contexts in which they are applied. In Section 3.8, we provide an analysis of the collected works, highlighting insights and the current limitations we have detected.

The content of this Chapter has been published in [12].

3.1 Introduction

Rapid technological advances and the amount of data have made AI a necessary tool in modern industry and research [23, 121, 203, 266]. Food engineering represents a perfect application for AI technology, as food requires in-depth study, processing, and analysis. The large volume of data generated in this field makes AI especially valuable for data analysis. However, the extensive use of AI introduces new questions about its trustworthiness and reliability.

To ensure trust in the results, it is necessary not only to understand the decision-making process behind the AI model but also to enhance its transparency, auditability, and informativeness [145]. Despite this, interpretable AI methods are still not widely adopted in the food sector, highlighting the need for greater focus on transparency and model explainability in this field. In response to this need, XAI has emerged as an important area of research to increase the trustworthiness of AI model predictions. It encompasses techniques aimed at elucidating the behaviour of these models by providing insights into their complex operations. In food engineering, XAI has been applied to allow accurate identification and validation of critical characteristics in tasks such as contaminant detection, nutritional value estimation, and product authentication, ensuring safety, transparency, and reliability in food quality control. This enables greater confidence by model users and customers, identifies

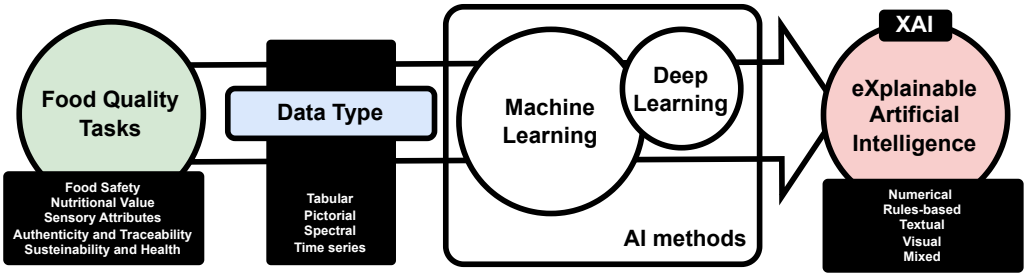


Figure 3.1: Overview scheme, from food quality tasks to XAI techniques. XAI is applied as an endpoint of a data processing pipeline that takes into consideration the task, type of data, and the specific AI model employed, such as ML and DL. According to these factors, one or more specific XAI techniques are employed, which produce explanations—tokens of information useful for model developers or users to gain insights into the prediction dynamics. Explanations can be produced in different types, each conveying a different facet of the information provided.

potential biases to improve accuracy, and supports the development of new, safer, and better-quality products.

Given the necessary role of food in human life, the food industry is keenly interested in applying these techniques to ensure the reliability of AI-driven outcomes [159]. However, we have identified several gaps in the literature linking XAI with food engineering. Firstly, there is a lack of standardization in the terminology and keywords used across various publications, creating challenges for data analysts and food engineers to communicate effectively. For instance, terms like "interpretation", "explanation", and "comprehension" are often used interchangeably for similar tasks, particularly when leveraging AI models in food quality research. Moreover, there is no comprehensive overview of the current state of the art that addresses these differences and provides insights into the advantages and drawbacks, which could be important for helping non-experts understand the progress and potential of these disciplines in research.

This literature review provides valuable insights for food industry specialists on the potential and importance of XAI. In particular, it offers an overview of current XAI applications in the food industry across key quality tasks—such as food safety, nutritional value determination, sensory attributes, authenticity and traceability, as well as sustainability and healthiness. We categorize applications by data type (tabular, pictorial, spectral, and time series) and forms of explanations generated by the applied XAI methods (numerical, rule-based, textual, visual, and mixed), highlighting its potential for further development, as depicted in Figure 3.1.

Additionally, our goal is to bridge the gap between the domains of XAI and food quality by presenting a taxonomy and arranging the current state of XAI applications in food research into an organized structure. Specific objectives are:

- to make a comprehensive survey and define a classification system to organize XAI methods applied to food quality;
- to introduce a taxonomy related to food quality to enhance understanding of the analyzed works;

- to summarize the XAI techniques used, detailing the types of data and AI methods employed in these studies;
- to offer an overview of the current state of XAI applications in the food sector, drawing insights from over a hundred papers;
- to provide comparative insights from the analyzed works, presenting intuitive connections between food quality tasks, data types, and XAI methods;

Only articles specifically addressing the topic of food quality were considered for this review. Papers that did not explicitly describe the use of a specific "XAI technique" in the field of food quality, even if they covered both food engineering and XAI, were excluded from our analysis. We performed an exhaustive search on *Google Scholar* and *Scopus* using the following keywords: "explainable artificial intelligence", "XAI", "food", "food science", "food quality", "food control", and "agriculture". These terms were strategically combined to cover relevant literature published over the past ten years. We examined the reference sections of the articles obtained in the initial search to identify additional relevant articles and integrated them into our research base. Lastly, we focused on several widely used XAI techniques, including LIME, SHAP, CAM, Grad-CAM, PDP, and LRP, introduced in Section 2.3.2. We investigated papers citing these works to identify any additional relevant articles and incorporated them into our research base. We did not record the exact number of papers found by entering the keywords in the mentioned search engines, but all papers containing these words were read and included if they met the inclusion requirements for the literature.

3.2 Taxonomy and Categorization Criteria

This Chapter presents the corpus of papers reviewed and analyzed for the literature review. The works are organized to help the reader locate existing explainability techniques, understand their application context, and quickly grasp what each method entails.

The categorization follows three criteria:

- **General taxonomy**. Introduced in this thesis and detailed in Section 3.2.1. It defines application areas within food quality analysis. Each subsequent section gathers works that fall under a taxonomy topic.
- **Type of explanation**. Defined using the classification described in Section 2.3.1. Each subsection groups works that employ techniques with the same type of explanation.
- **Data type**. Drawn from the categories outlined in Section 3.2.2. Each paragraph within a subsection presents works that use the assigned data type.

To improve readability, the Chapter also offers a brief overview of the primary AI methods used in the collected papers and their relationship to explainability in Section 3.2.3. Because of space constraints, this overview is necessarily selective and does not claim to be exhaustive.

3.2.1 Overview of the Proposed Taxonomy

For a comprehensive analysis of food quality, we propose a taxonomy encompassing five main topics: food safety, nutritional composition, sensory attributes, authenticity and traceability across the supply chain, and sustainability and health within the context of food engineering and nutrition. Each of these topics offers a detailed understanding of the elements and challenges that comprise food quality, reflecting consumer needs and expectations [207].

Food Safety: Food safety involves the assurance that food is free from agents that may pose a health risk. In addition to implementing rigorous hygiene procedures and sanitary practices to minimize contamination risks, controlling pathogens such as bacteria, viruses, and parasites is fundamental. Furthermore, the presence of pesticide residues, heavy metals, and harmful chemical additives must also be strictly controlled. Specific regulations limit the concentration of these contaminants in food to ensure consumer safety [95].

Nutritional Value: Nutritional value is directly related to food composition and how it impacts human health and well-being. Foods rich in vitamins, minerals, proteins, carbohydrates, and healthy fats are necessary for the proper functioning of the body and prevent nutritional deficiencies based on their compounds. Besides nutritional content, the bioavailability of nutrients is an important quality aspect of food [233].

Sensory Attributes: The sensory requirements of food are directly perceived by consumers, making them a necessary means of interaction between products and consumers. Attributes like colour, shape, and taste, along with other appearance attributes, are key indicators of quality and freshness. Sensory standards are relevant for denoting fresh food, which usually has higher nutritional value and consumer acceptability [157].

Authenticity and Traceability: The authenticity and traceability of food ensure compliance with legal standards and increase consumer confidence. Identifying and preventing fraudulent practices, such as food adulteration and counterfeiting, is necessary to guarantee product authenticity. They not only indicate authenticity but also verify species variety and monitor environmental conditions during cultivation, production, and storage, thereby ensuring food quality and sustainability [56, 60, 276, 277].

Sustainability and Health: Sustainability and health are important for the availability of food with desirable sensory and physicochemical characteristics while also guaranteeing animal welfare, environmental preservation, and consumer health. The use of technologies to analyze phenotypic characteristics of plants has promoted more resilient and nutritious crops. The implementation of automated processes in food production increases efficiency, reduces waste, and improves food safety [28, 162]. We differentiate health from nutritional value by defining it more broadly to include disease prevention, immune support, mental health, and the effects of food processing, additives, and potential allergens.

3.2.2 Classification Method Based on Data Types

With the continuous advancement of technology, food quality analysis has significantly evolved, leveraging the diversity of sensors, methods, and devices to collect data into datasets. These datasets encompass various modalities, including tabular data, images or pictorial data, spectral data, and time series data, each offering distinct advantages for analysts in evaluating relevant aspects of food quality. The complexity and volume of these data have necessitated AI to process large datasets automatically, identify complex patterns, and extract the maximum useful information from this diverse data.

Tabular data: Tabular data allow for systematic and clear organization of information, which can simplify statistical analyses and data management. However, complexity can arise from integrating interrelated variables. By using AI algorithms, it is possible to explore these datasets to identify non-obvious correlations and interactions between variables, enabling advanced predictive analyses.

Pictorial data: Pictorial data allows for clear and intuitive visualization of information, facilitating the communication and understanding of complex data. They enable the identification of small defects or imperfections in food, such as stains or deformities. Additionally, the images are the results of several non-destructive techniques that support sustainable analysis and monitoring without the need for chemical reagents required in other conversion techniques. Pictorial data include Hyperspectral Imaging (HSI), *X-ray imaging*, and *multispectral imaging*, all of which are widely applied in the food quality sector.

Spectral data: Spectral data allow for detailed and precise analysis of chemical interactions through the analysis of electromagnetic radiation emitted, reflected, or absorbed at different wavelengths. This makes spectral data a highly accurate tool for detecting small changes in the composition of the analyzed food, providing insights that conventional methods may not reveal. Like pictorial data, spectral data are obtained through "green", non-destructive techniques. Methods such as Near-InfraRed (NIR) and *Raman* spectroscopy, along with Proton Nuclear Magnetic Resonance (^{1}H NMR), offer high precision comparable to imaging techniques but at a lower computation cost.

Time series data: Time series data enables continuous and dynamic monitoring of various factors over time. These data capture temporal variations in critical parameters, providing detailed insights into trends and anomalies that may arise at different stages of the production and distribution chain. Additionally, environmental sensors use sequential measurements to establish reference parameters over time.

3.2.3 Artificial Intelligence Methods in the Reviewed Studies

With access to a wide array of pre-built libraries and proven techniques, researchers can adapt various AI methods to address their specific challenges. As access to data increases, data analysts, including chemometricians, can utilize AI methods and improved resources to apply their techniques more effectively. This capability enables them to discover more efficient and straightforward solutions that are specifically tailored to their data and the goals they wish to achieve.

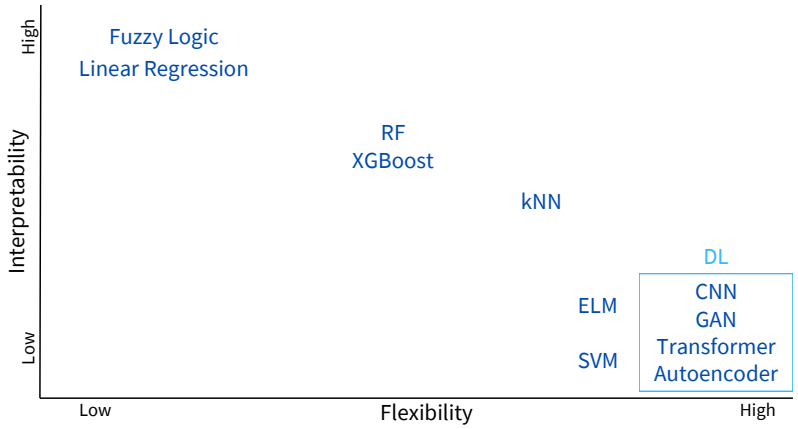


Figure 3.2: The trade-off between expressivity or flexibility and interpretability of the AI models exploited in the reviewed studies. Expressive models, such as those based on DL, are capable of reaching higher task-level performance but are often difficult to interpret. On the other hand, less complex models, like LR, are inherently interpretable, but often incapable of attaining high task-level performance.

Among the works analyzed, only a few propose using classic AI algorithms, such as *Fuzzy Logic* [230, 268]. While these algorithms offer the advantage of transparency due to their reliance on well-defined rules, they also demand a deep understanding of the problem and the precise formulation of logical frameworks.

A significant portion of the analyzed articles focuses on using ML methods. LR algorithms are commonly employed for their effectiveness, simplicity, and complete transparency [54, 232]. Similarly, ensemble methods such as RF [249, 273] and XGBoost [111, 219] are widely favoured for their robustness to outliers and their ability to capture intricate relationships within the data. Although they are generally straightforward to explain, their complexity increases as the number of base learners grows. Some studies utilize unsupervised ML techniques, such as kNN [57] and *Clustering* [67, 227], which offer transparent and relatively interpretable decision-making processes. Support Vector Machine (SVM) are also commonly used techniques [262, 311], although their decision-making process is more complex and harder to interpret. Extreme Learning Machine (ELM) [173] is also noted for its fast learning speed, though it can be challenging to interpret.

Most of the analyzed works leverage DL techniques due to their ability to learn complex patterns and extract valuable information from highly intricate data, such as images. NNs can achieve outstanding performance even on highly complex problems; however, this comes at the expense of being extremely difficult to interpret. CNNs are the most widely used methods for image analysis because of their effectiveness in generalizing and extracting meaningful features. By modifying their architecture—such as internal layers or final classifiers—more specialized networks can be developed, such as VGG or Residual Network (ResNet) for greater robustness [72], *MobileNet* or *EfficienNet* for lightweight applications [43], or YOLO model for object detection [20]. Additionally, other DL architectures are employed, including attention-based models like ViT [166], as well as models that generate or synthesize data, such as GAN [282] and *Autoencoders* [271], are also employed. These models,

when combined with other techniques or used for feature extraction, can significantly enhance performance. However, while their outputs are often understandable, the models themselves remain difficult to interpret.

Figure 3.2 presents an overview of the relationship between interpretability and flexibility, as discussed in Section 2.3, for the AI methods used in the reviewed studies.

3.3 Explaining Food Safety

We observe that most of the applications of XAI techniques in the field of Food Safety focus on providing visual explanations, as depicted in Table 3.1. The reason is the frequent use of pictorial data by researchers to study diseases affecting food and insects attacking plants. The images are typically processed using CNNs, with CAM [308] and derived XAI techniques widely applied to explain them.

Table 3.1: Summary of the works introducing applying XAI for food safety surveyed in Section 3.3, according to their data type and explanation type (labelled as "Expl. type").

Works	**Data type**	**Expl. type**
[4, 5, 15, 16, 20, 22, 25, 33, 42, 44, 45, 46, 51, 53, 72, 73, 82, 83, 90, 98, 105, 117, 123, 124, 128, 133, 147, 165, 166, 167, 170, 173, 183, 189, 192, 194, 195, 198, 204, 210, 211, 214, 218, 231, 238, 246, 247, 261, 272, 279, 283, 292, 301, 307, 309]	Pictorial	Visual
[284]	Spectral	Visual
[43, 48, 50, 168, 296]	Pictorial	Mixed
[232]	Tabular	Mixed

3.3.1 Visual Explanation

Numerous studies using **pictorial data** have focused on detecting plant diseases in staple crops like maize, rice, and wheat, showcasing the importance of accurate disease identification in the food supply. [42] developed a transfer learning methodology enhancing MobileNetV2 with CAM to diagnose plant diseases in maize and rice. [46] and [98] used CNN models to detect maize and peanut diseases, applying channel attention and pruning techniques for improved feature extraction. Similarly, [72] applied CNNs with Grad-CAM to distinguish healthy from infected wheat, effectively identifying disease-affected areas. [170] introduced C-DenseNet, a modified CNN model, to grade wheat stripe rust severity, validated using Grad-CAM++ [41]. [82] developed a YOLOv5s-based model with MobileNetV3 and C3Ghost modules to detect Fusarium Head Blight (FHB) in wheat, using Grad-CAM. In addition, addressing the complexities of varying disease images, [43] improved MobileNetV2 with

a Location-wise Soft Attention mechanism and CAM, demonstrating its practical utility in identifying crop diseases in diverse conditions.

In contrast, some studies explored alternative crops and agricultural sectors. [195] used high-resolution video data and a CNN-based object detection model to monitor pecan tree health, focusing on *xylella* disease, validated by Grad-CAM to highlight critical canopy features. [307] introduced T-RNet, a Transformer-Embedded ResNet model, for cassava leaf disease detection, demonstrating focus on relevant areas through Grad-CAM visualizations.

Other studies have focused on tomato and potato disease detection, employing diverse models to improve early and accurate identification. [51] developed an EfficientNet-based model to classify tomato diseases from segmented leaf images, using ScoreCAM [280] for early detection validation. Similarly, [246] combined InceptionNet and U-Net, two CNNs, for tomato disease detection and segmentation, validated by ScoreCAM. [194] proposed an ensemble model combining DenseNetMini with Gradient Product optimization and Grad-CAM to enable interpretable disease detection in plant leaves, specifically for tomato and apple plants. [124] introduced DVTXAI, a Deep ViT model integrated with SHAP, for identifying infections in tomato and potato plants. [5] developed ExE-Net, an Explainable Ensemble Network for potato leaf disease classification, integrating various CNN-based models with XAI techniques—including LIME, SHAP, and Grad-CAM—to enhance the accuracy and interpretability of potato disease identification.

Similarly, [90] applied LIME and Grad-CAM to a DenseNet-based model developed for classifying tomato leaf diseases. Moreover, [117] proposed the use of a MobileNetV2 model combined with data augmentation and reweighting techniques for accurate classification of potato leaf diseases on imbalanced datasets, with Grad-CAM used to explain the model's predictions. [22] introduced a novel saliency-based XAI method using perturbation techniques for object detection, which iteratively refines saliency maps to enhance the interpretability of the applied ResNet model while maintaining high accuracy in classifying potato diseases. Finally, [214] presented a tomato health monitoring system that integrates YOLOv8 for detection and MobileNetV3 for real-time counting and classification of diseases, with Grad-CAM++ used to explain the model's predictions. It is important to mention that CAM-based technology contributed to model verification and highlighted regions with particular texture and color patterns.

A set of research targeted tree and fruit diseases. [165] applied CNN models to detect grape diseases using the PlantVillage dataset, validated with Grad-CAM. [183] introduced GLD-Det, a MobileNet-based model for detecting guava leaf diseases, confirmed by Grad-CAM for real-time mobile applications. This is an important example of a CAM-based solution in real-time prediction, improving the sample prediction explanation. [283] explored the interpretability of CNN models like VGG, GoogLeNet, and ResNet for fruit leaf classification, demonstrating the superior performance of ResNet with Grad-CAM for feature visualization. In another example, [301] integrated a novel module into CNN architectures for fine-grained crop disease classification, with Grad-CAM confirming the model's focus on relevant features.

In addition to disease detection, several studies shifted attention toward pest detection and pest management in agriculture. [33] evaluated Faster-RCNN with

a MobileNetV3 backbone for pest identification, validated with Grad-CAM. [16] improved pest classification models using genetic algorithms, confirming model efficiency through Grad-CAM visualizations. [309] developed ExquisiteNet, a DL model for pest identification validated by Grad-CAM. Additionally, [53] used various XAI techniques to provide detailed visual explanations for a lightweight CNN in crop health monitoring. [173] utilized LIME with the proposed I-LDD framework, leveraging ELM for fast and robust disease classification on the PlantVillage dataset, accompanied by visual explanations that highlight diseased leaf areas. These two papers show composite solutions, utilizing multiple XAI techniques to provide more insightful explanations.

Some studies extended the use of XAI techniques beyond agriculture. [167] applied Grad-CAM to validate an ML method for classifying mercury exposure in fish, supporting food safety beyond agriculture. [83] introduced EffiNet-TS, a model based on EfficientNetV2, incorporating an NN to reconstruct images that highlight key symptoms, thereby clarifying the decision-making process. [211] proposed a customized EfficientNetB4 model for high-precision classification of chill leaf diseases, validating the model using Grad-CAM. Similarly, [238] evaluated the performance of four CNN models, with EfficientNetB4 performing best on a dataset of diseased and healthy plant leaves, confirming the models' focus on critical disease features like rust pustules through Grad-CAM. [45] introduced a meta-learning approach for plant disease detection, interpreted with Task Activation Mapping, a CAM-based technique specifically developed for this study. [44] developed a convolutional ensemble network using lightweight CNNs like MobileNetV2, validated by Grad-CAM. Furthermore, [166] employed a ViT model for plant disease classification, with Grad-CAM confirming the model's focus on relevant disease features.

Recent research has significantly advanced the use of DL models for mobile device deployment in agricultural disease detection. [20] adapted a YOLOv8n model for real-time wheat ear detection, optimized for mobile devices. [4] utilized MobileNetV2 in detecting tomato leaf diseases, emphasizing its suitability for low-end devices in real-world applications. [25] proposed the CD-MobileNetV3 model for identifying corn leaf diseases, demonstrating its efficiency for mobile use. Likewise, [147] applied the lightweight ShuffleNetV2 model to detect corn seed diseases. These studies validate their models using Grad-CAM for real-time deployment on mobile platforms, highlighting the increasing role of mobile-optimized models in advancing agricultural monitoring and management.

There are numerous applications of AI in the field of fruit quality, particularly those involving the use of XAI to enhance trust in model predictions. An important case is presented by [198], who proposed an ensemble learning framework for fruit plant disease detection using multiple DL models, incorporating LIME across all models as an additional tool for result evaluation. [15] proposed a method for classifying various banana diseases—including Cordana, Black Sigatoka, Pestalotiopsis, and Fusarium Wilt—by analyzing leaf images using EfficientNetB0 and employing Grad-CAM to enhance classification accuracy and interpretability. [105] introduced an interpretable AI-based method for localizing mildew symptoms in grapevine using EfficientNetV2S and Grad-CAM. [210] presented LEViT, a ViT model for tree leaf disease classification, incorporating Grad-CAM to ensure reliable and interpretable

results. An example highlighting the need to apply multiple XAI techniques for reliable results is [128], who developed an AI-based system for date palm classification—capable of identifying diseases and assessing fruit ripeness—using VGG16 in combination with SHAP, LIME, Grad-CAM, and Grad-CAM++. [247] proposed a modified MobileNetV2-based model to enhance the classification of cucumber leaf diseases, ensuring result explainability through the integration of LIME. [189] aimed to improve the explainability of DL models—specifically a ResNet50 model—used in citrus disease detection, by introducing a novel model-agnostic, local explainer for image-based classification called the Multi-objective Genetic Algorithm Explainer. [231] introduced a CNN-based approach for detecting mulberry leaf diseases, utilizing the MobileNetV3Small model and Grad-CAM to align model predictions with expert assessments.

More recent applications focus on crops, the primary source of human sustenance, highlighting the growing role of AI and XAI in ensuring food security [133] developed MaizeNet, a CNN framework combining clustering for maize crop image segmentation and classification. Grad-CAM was applied to explain the model, providing severity assessments and crop loss estimation. [279] employed CNNs to quantify rice grain chalkiness caused by high nighttime temperatures, using Grad-CAM to localize affected areas. [292] proposed a convolution-based method for rice disease detection, with Grad-CAM highlighting the model's effectiveness even in complex scenarios. [73] proposed a novel DL model that combines DenseNet for feature extraction with an SVM for classifying healthy and diseased sugarcane plants, incorporating LIME to enhance trust and usability. [261] and [204] applied LRP to enhance VGG16 models for identifying crop leaf diseases, aiming to improve performance. [218] developed a deep transfer learning-based framework for diagnosing rice leaf diseases, leveraging various DL models and integrating Grad-CAM to enhance the system's reliability for farmers. [192] incorporated LIME into an EfficientNet-based model to address trust issues in plant disease classification. Since Maize Streak Disease poses a serious threat to maize crops, [123] introduced a CNN-based framework for its diagnosis, incorporating SHAP and LIME. Finally, [272] proposed a comparative framework integrating Bayesian optimization for hyperparameter tuning across CNN-based models—InceptionNet, MobileNet, ResNet, and RegNet—to diagnose rice plant diseases, leveraging LIME to enhance the interpretability of model behaviour.

Considering **spectral data**, [284] developed a method using HSI and DL to assess FHB infection levels in wheat kernels, extracting reflectance spectra and selecting optimal wavelengths. A residual attention CNN classified infection degrees, distinguishing features across infection levels, as confirmed by Grad-CAM. Although spectral data is key for food safety, it does not significantly use visual explanations.

3.3.2 Mixed Explanation

Several studies have proposed DL methods to address food safety issues using **pictorial data**. However, they applied different XAI techniques than those previously discussed, resulting in distinct explanation types.

Recent advancements in DL have focused on enhancing food safety by employing

various XAI techniques to provide insights into model decisions. [58] explored the application of CNNs for plant disease diagnosis, utilizing XAI methods like LIME, Grad-CAM, and SHAP to offer both visual and mixed explanations. [296] introduced a novel workflow using ResNet18 for pest recognition, which involved segmenting images into meaningful concepts and explaining decisions through weighted directed graphs and concept importance, improving transparency but noting the complexity of explanation generation. [48] combined DL with semantic web technologies for cassava disease detection, utilizing a ViT and a semantic model that integrates environmental data. This approach achieved high accuracy and introduced a unique explainability method using knowledge graphs tailored for end users. [50] proposed using both visual and numerical explanations from LIME to provide localized FI, enhancing the transparency of a CNN-based model for classifying rice crop diseases. [168] presented PLD-Det, an improved YOLOv7-based real-time plant leaf disease detection model, incorporating SHAP explanations to enhance transparency and make predictions more interpretable for farmers.

Regarding **tabular data**, [232] introduces a novel model for classifying pistachio species by combining feature selection, XAI-based interpretation with LIME, and classification with LR with 90.0%.

3.4 Explaining Authenticity and Traceability

By addressing the authenticity and traceability of the food supply chain, we identified a wider application of XAI techniques. This area emerged as the second most significant food-related task application of XAI. Table 3.2 summarizes the works surveyed in this section.

Table 3.2: Summary of the works introducing applying XAI for authenticity and traceability surveyed in Section 3.4, according to their data type and explanation type (labelled as "Expl. type").

Works	**Data type**	**Expl. type**
[110, 130, 146, 191, 193, 220, 264, 290, 291, 295, 305]	Pictorial	Visual
[299]	Pictorial	Mixed
[118]	Spectral	Visual
[286, 289]	Tabular	Visual
[3, 19, 24, 40, 71, 78, 104, 111, 131, 132, 139, 153, 156, 164, 175, 184, 213, 235, 241, 267, 294, 298, 311]	Tabular	Numerical
[268]	Time series	Numerical
[34, 38, 70, 109, 137, 208, 219, 229, 249, 273]	Tabular	Mixed
[282]	Time series	Mixed
[108]	Spectral	Mixed

3.4.1 Visual Explanation

Recent studies using **pictorial data** have advanced the variety traceability and authenticity verification of agricultural products. [305] developed a CNN model for herb variability identification, using Grad-CAM to highlight relevant herb parts while ignoring background noise. [290] focused on maize seed classification with a ResNet model, while [295] applied HSI and DL to classify hybrid okra variability seeds. More recently, [220] proposed the application of various CNN models to classify fungal species, followed by the use of Grad-CAM to interpret the model predictions

Beyond traceability, several studies addressed the identification of damaged and adulterated products. [146] used a ResNet18 model to detect cocoa beans with bad fermentation, with Grad-CAM providing interpretability. [110] developed a lightweight CNN, the Soybean Network, to classify damaged soybean seeds, enhancing quality inspection through Grad-CAM visualizations. Meanwhile, [191] introduced CondimentNet, an optimized ResNet18 model, leveraging Grad-CAM to detect adulteration in various condiments.

[264] and [130] emphasized improving agricultural and food production processes for quality and sustainability. [264] developed BraeNet, a modified ResNet classifier using 2D and 3D X-ray imaging to detect internal browning in Braeburn apples, demonstrating the practical application of radiography in inline quality sorting. Similarly, [130] explored food supply chain optimization, covering plant growth prediction, energy-efficient refrigeration, and expiry date recognition, reinforcing the role of process improvements in maintaining food quality and safety.

[193] and [291] proposed the use of Unmanned Aerial Vehicle (UAV) aerial imagery as the primary pictorial data source for two similar AI-based applications. [193] explores an interpretable AI-based approach for identifying and mapping weeds and crops using UAV imagery, applying U-Net for segmentation to filter noise and extract key regions, followed by ViT for classification. XAI techniques such as LRP and Pixel Density Analysis are employed in the classification process to enhance transparency. [291] investigated optimal input image conditions for rice yield prediction using CNN models applied to UAV aerial images captured after the mid-ripening stage, assessing the results with XAI techniques such as Gradient-Based Feature Importance Analysis.

[118] proposed using **spectral data** to address a traceability problem by developing a rapid, non-destructive method for identifying counterfeited beef adulterated with colourants and curing agents. Applying Grad-CAM to spectral data improved the method by generating visual explanations that highlighted key wavelengths influencing the model's decisions.

Using **tabular data**, [286] highlighted the importance of accurate crop yield forecasting in addressing food quality challenges arising from climate change, population growth, soil erosion, and decreasing water resources. The regression model achieved good performance with activation maps to visualize and analyze the features driving the yield predictions, demonstrating that the length of the growing season and conditions such as temperature and sunlight were critical factors. Similarly, [289] presented an ML framework for agricultural drought prediction in the Tapieh Mountains, China, including SHAP analysis to visually highlight the most

influential meteorological factors contributing to drought severity.

3.4.2 Numerical Explanation

Several studies have applied advanced ML techniques using **tabular data** for crop yield prediction, integrating multiple data sources and employing SHAP for interpretability. [267] demonstrated the effectiveness of using SHAP with an AI model for tabular data analysis in aeroponics through data fusion from multiple sensors. Similarly, [3] used XGBoost and SVM to analyze factors affecting rice production, validating model decisions with LIME. [111] applied XGBoost for soybean yield prediction, with SHAP highlighting key factors such as near-infrared light and temperature. [139] further explored soybean yield estimation, emphasizing the role of the vegetation index using SHAP.

Some studies incorporated satellite and meteorological data for improved predictions. [164] utilized LSTM trained on multisource data, applying Integrated Gradients and SHAP to identify critical factors like enhanced vegetation index and temperature. [104] examined the impact of extreme weather on crop yields, revealing sensitivity differences among crops and regions.

Soil water content has also been a focus of ML models in agricultural management. [294] introduced Tree-structured Parzen Estimator (TPE)-Categorical Boosting algorithm (CATBoost), incorporating soil moisture and environmental factors, with SHAP demonstrating model sensitivity to environmental changes. [298] used TPE-GBDT to map soil water content across the Yellow River Delta, identifying key variables such as soil texture and vegetation. [311] applied SVMs and SHAP to highlight necessary factors in digital soil mapping, reinforcing the integration of terrain and geological data for effective agricultural management. The idea of selecting the most suitable soil has also been explored by [24], who aimed to improve crop quality by classifying different soil types using an ML model, and applied SHAP to highlight the most important features influencing the model's decisions. Similarly, [40] presented an RF model for predicting soil fertility, using SHAP to highlight various physicochemical soil properties that determine fertility levels.

There is also a substantial body of work focused on the traceability and analysis of environmental conditions to enhance the production and quality of crops such as rice, wheat, and maize, leveraging SHAP or LIME to identify the most influential features utilized by the models in performing the given tasks. [153] proposed an ML model for crop prediction, integrating Genetic Algorithms for hyperparameter optimization and RF for classification, while applying XAI techniques such as LIME and SHAP to enhance classifier interpretability—ultimately supporting farmers in optimizing agricultural planning, reducing crop losses, and improving productivity. [19] presented ML models for crop classification and yield prediction, leveraging XAI techniques such as LIME and FI to enhance model interpretability. Similarly, [156] aimed to provide accurate crop yield predictions by using generative algorithms to optimize a Deep Neural Network (DNN), and employed LIME to explain the model's outputs. [184] proposed a method for selecting optimal crops based on environmental and soil conditions, utilizing Radial Basis Functions and SHAP. [241] introduced XAI-CROP, an ML-based crop recommendation system improved by

including LIME to explain predictions, designed to assist farmers in selecting optimal crops by analyzing soil characteristics, historical crop performance, and weather patterns. A similar tool was developed by [132], who employed various ML models to recommend optimal crops for specific regions, analyzing the results using LIME and SHAP. [131] aimed to enhance the interpretability of AI-driven crop yield predictions by integrating saliency maps and SHAP analysis into kNN models. [78] leveraged an XGBoost model combined with SHAP values to map and understand the influence of weather and soil variables on wheat yield in Eastern Australia. [175] introduced ML-based regression methods along with XAI techniques—SHAP and LIME—to predict crop yields and assess the impact of climate change on agriculture.

Finally, several studies propose applications similar to those previously discussed, but adapted to different food products. In particular, [213] applied ML models—specifically tree-based ensemble methods—and LIME to classify blackcurrant powders based on image texture features. [235] proposed using ML-based models, such as RF and SHAP, to enhance coffee quality assessment—traditionally reliant on subjective evaluation—by contributing to the standardization of coffee grading. [71] examined the integration of ANN and XAI techniques, such as FI, to enhance quality control strategies in the agri-food industry, with a specific focus on milk quality classification.

3.4.3 Rule-based Explanation

Authenticity and traceability have not been deeply explored in rule-based explanations. [268] highlighted the need to monitor low-cost, automated, and interpretable irrigation systems using time series data. To address this, they proposed a new system called Vital, which integrates Internet of Things (IoT) sensors, a data management platform, and a fuzzy rule-based decision support system to automate irrigation. The system was evaluated through pilot cases and effectively automated the irrigation process, monitoring and managing open-field installations that provided water.

3.4.4 Mixed Explanation

In [299], various XAI techniques were applied to enhance the authenticity verification of honey products using HSI, addressing challenges related to high dimensionality and noise through the use of **pictorial data**. By integrating multiple XAI algorithms with CNNs, they developed a wavelength selection method to identify the most informative spectral bands, effectively reducing data dimensionality, particularly in classifying honey by botanical origins.

Tabular data was explored by [273] and [229] applied various XAI techniques to enhance the interpretability of ML models in agricultural analysis. [273] developed an RF model to assess the influence of biophysical, bioclimatic, and socioeconomic factors on land use for wheat, maize, and olive groves, with FI, PDP, and LIME identifying key variables such as drainage density, slope, and soil type. Similarly, [229] investigated the effects of no-tillage on maize yield using ML and XAI methods, pinpointing critical biophysical and climatic factors. [219] and [137] demonstrated how XAI techniques, when integrated with ML, provide insights into agricultural

expansion and product quality assessment. [219] applied XGBoost and SHAP to analyze avocado frontier expansion, visualizing key environmental and accessibility factors. [137] used XGBoost with SHAP and PDP to evaluate liquor quality in the Vinho Verde region, identifying key chemical attributes influencing product quality. [249] utilized an RF model with LIME to examine the long-term impact of climate variables and soil properties on crop yields in the Coterminous United States. The study identified critical environmental factors affecting yields, demonstrating the value of XAI for understanding complex agricultural data and supporting climate adaptation strategies for stakeholders. [70] employed XGBoost and SHAP to predict annual palm oil yield in Indonesia by analyzing fifteen agrometeorological variables, including rainfall rates, number of rainy days, and soil properties. [109] proposed a Bayesian ensemble model to analyze the impact of climate on crop yields, effectively separating climate and technological influences while capturing nonlinear climate effects, resulting in high accuracy and interpretable outcomes. [208] explored the application of XAI techniques—specifically LIME and SHAP—to enhance the transparency and user understanding of ML-based models applied to agricultural tabular data, focusing on two case studies: wheat yield prediction and grape yield prediction for wine production. [34] demonstrated that XAI techniques can enhance transparency in food fraud detection by applying LIME, SHAP, and the What-If Tool [285] to DL models. Finally, [38] proposed the application of various ML-based models, including LR, CATBoost, kNN, and RF, for automated rice classification in Cammeo and Osmancik rice species. To ensure transparency, SHAP and Individual Conditional Expectation plots [85] were employed.

Conversely, using **spectral data**, [108] investigated ^{1}H NMR spectra to determine the geographical origins of Asian red pepper powders, employing ML, SVM, and CNN models with dimensionality reduction techniques. Grad-CAM and SHAP provided insights into the decision-making processes, highlighting metabolite distribution variations as key classification factors. This study demonstrated the potential of these models for broader applications in food authenticity verification.

Time series data was also explored; for example, [282] introduced DeepFarm, a DL framework for managing and predicting agricultural production under uncertainties such as natural disasters and cyber-attacks. Using DL and causal inference, DeepFarm accurately predicted crop yields across U.S. regions, with precipitation anomalies notably impacting corn yields.

3.5 Explaining Nutritional Value

Studies on nutritional property explanations reveal a predominant reliance on visual explanations using pictorial data, with minimal use of rule-based methods and occasional mixed explanation types. Table 3.3 summarizes the studies surveyed in the present section.

3.5.1 Visual Explanation

Several studies, using **pictorial data**, have leveraged DL models and XAI techniques to enhance food classification and nutrient estimation. [281] applied a weakly

Table 3.3: Summary of the works introducing applying XAI for nutritional value surveyed in Section 3.5, according to their data type and explanation type (labelled as "Expl. type").

Works	Data type	Expl. type
[97, 114, 116, 140, 141, 149, 150, 151, 161, 169, 174, 197, 199, 242, 243, 257, 258, 281, 303]	Pictorial	Visual
[122]	Spectral	Visual
[57, 64, 67, 92, 127]	Tabular	Numerical
[227]	Pictorial	Rule-based
[94]	Tabular	Rule-based
[54]	Tabular	Mixed

supervised VGG16-based CNN for food image segmentation, using Instance Activation Maps to highlight relevant regions. [161] introduced the Wide-Slice Residual Network, incorporating slice convolution blocks for improved nutritional evaluation through Grad-CAM visualizations. [174] estimated vegetable mass using CNNs and monocular RGB images, while [197] utilized attention mechanisms for classifying unlabeled food images from social media.

Some works focused on user-centric approaches for food recommendation and recognition. [303] introduced JDNet, a CNN-based model for mobile food recognition, validated through Instance Activation Maps. [116] used Grad-CAM to enhance ingredient recognition in a few-shot learning framework, while [169] developed PiNet, a multi-task learning framework improving food recommendation by integrating visual and semantic features.

Optimizing food recognition for edge devices has also been explored. [257] developed a MobileNetV3-based system, incorporating a user-centered XAI framework with Grad-CAM++ for dietary assessments. [151] proposed a big data-driven approach for nutrient estimation, visualizing critical regions with Grad-CAM. [199] applied ResNet34 to predict the mechanical properties of Granny Smith apples, using Grad-CAM saliency mappings to reveal biophysical tissue changes.

[150], [141], and [258] contributed to dietary assessment and food image recognition. [150] introduced the ChinaFood-100 database, evaluating multiple DL architectures and using Grad-CAM to validate nutrient predictions. [141] explored oriental food recognition with VGG16 and InceptionNet, revealing model inconsistencies through LIME and Grad-CAM. [258] developed a dietary assessment system combining ELM with a SHAP-guided feature selection strategy.

Beyond classification, some studies integrated advanced DL architectures for food analysis. [242], [114], and [149] developed non-destructive evaluation and ingredient prediction models. [242] proposed the Swin-Nutrition model, a transformer-based framework validated with Grad-CAM. [114] used EfficientNetV1 for allergy prediction and food classification, highlighting critical features with Grad-CAM. [149] introduced CACLNet, improving ingredient prediction by addressing class imbalance and background noise through Grad-CAM visualizations.

A multi-modal approach has also been explored to enhance nutrition estimation and food recognition. [140], [97], and [243] combined diverse data types and

learning techniques. [243] improved nutrition estimation using ResNet101, integrating multiscale image and depth data features. [97] introduced DPF-Nutrition, a transformer-based approach that generates depth maps for enhanced nutrient estimation. [140] developed MVANet, a multi-view attention-based CNN incorporating ingredient and recipe semantics, validated with Grad-CAM for food recognition in healthcare applications.

3.5.2 Numerical Explanation

Spectral data was explored in [122]. The authors employed visible NIR point spectroscopy to estimate sugar content in grape varieties at different maturity stages. Regression ML algorithms and a CNN were applied, with XAI techniques such as Variable Importance in Projection and Gini Importance validating the models and identifying key spectral features. On the other hand, **tabular data**, was discussed in [57], [92], [67], and [127] apply ML techniques to various food-related challenges. [57] used XGBoost to estimate added sugar content in foods, with SHAP enhancing model transparency for consumer awareness in regions without mandatory labeling. [92] developed the Flavonoid Astringency Prediction Database, employing ML models like RF to explore the relationship between molecular structures and flavor properties. Similarly, [67] applied ML to differentiate pepper spices during storage, using SHAP to identify key organic compounds. [127] developed an XGBoost-based model for predicting drug-food interactions using molecular fingerprint similarities, with SHAP providing insights into influential features relevant to clinical applications and dietary planning. [64] proposed a graph-based ML approach to predict the outcomes of formulation trials, aiming to reduce laboratory experiments, material waste, and development time in food design. To enhance interpretability, they applied GNNExplainer [293], a global explanation method tailored for graph NNs.

3.5.3 Rule-based Explanation

Only two significant studies employed XAI techniques to generate rule-based explanations in the context of Nutritional Values. [227] exploited **pictorial data** to propose a similarity score based on user community preferences, enhancing recommendation quality. The rule-based explainability method assigned each image to an appropriate food diet based on user profiles, supporting personalized dietary recommendations. [94] presented a novel no-code methodology for developing predictive models to classify the antioxidant activity of phenolic compounds, leveraging Decision Tree-based algorithms and Conceptual Density Functional Theory (CDFT) descriptors. The resulting models achieved high accuracy and full explainability through explicit, interpretable if–then rules derived from molecular features.

3.5.4 Mixed Explanation

Tabular data was explored in [54], introducing the Taste Peptide Docking Machine, a computational framework for predicting umami and bitter tastes in peptides. The framework integrates ML algorithms with molecular representation schemes, including docking analysis, molecular descriptors, and molecular fingerprints. SHAP and

LIME were applied to enhance interpretability, providing insights into key molecular features influencing taste prediction.

3.6 Explaining Sensory Characteristics

Sensory characteristics are extremely important for quality control, leading to the widespread use of sensors designed to mimic human senses. Among these, spectral devices—commonly used and established in the industry—offer rich information, suggesting potential applications for XAI techniques. However, it was observed that most studies focus on **pictorial data** and **visual explanations**, with only two works to date addressing spectral data for explainability. In Table 3.4 we summarize the studies surveyed in the present section.

Table 3.4: Summary of the works introducing applying XAI for sensory characteristics surveyed in Section 3.6, according to their data type and explanation type (labeled as "Expl. type").

Works	Data type	Expl. type
[36, 84, 100, 134, 136, 152, 179, 239, 260, 263, 271, 288, 302]	Pictorial	Visual
[222, 245]	Spectral	Visual
[6, 68, 154]	Tabular	Numerical

3.6.1 Visual Explanation

The studies highlight advancements in fruit integrity assessment using DL models and XAI techniques over **pictorial data**. [271] employed X-ray radiography and DL methods, including autoencoders and CNNs, for deep anomaly detection of internal defects such as browning and cavities, with heatmaps enhancing interpretability. [136] introduced MBNet, a CNN-based model utilizing sensory data from multiple cameras for pear evaluation. [263] applied UNet with synthetic data for internal pear defect segmentation, validated through Grad-CAM heatmaps. [179] used DenseNet201 for fruit quality classification, with Grad-CAM confirming its focus on relevant features. [36] investigated bruise detection in plums using HSI and CNNs, with Grad-CAM visualizations validating model predictions.

Food freshness assessment has also benefited from DL and HSI. [302] developed a VGG16-based model to classify shrimp freshness from smartphone images, using Grad-CAM to confirm inference regions. Similarly, [152] employed a colourimetric sensor and RGB images to monitor salmon freshness, with Grad-CAM revealing that the CNN prioritized sensor data over visual texture, emphasizing odor's role in freshness detection.

Beyond fruit, cereal integrity has been explored using XAI methods. [134] applied Grad-CAM in an EfficientNet-B3-DAN model to detect rice germ integrity, confirming the model's focus on relevant features. [288] addressed crop yield estimation by developing an Inception-ResNet-based regression model for leaf counting,

handling occlusions in monocots. Grad-CAM analysis confirmed its focus on leaf tips, validating effectiveness across sorghum and maize datasets. [239] enhanced crop classification by employing a MobileNetV2 model validated with Grad-CAM to assess the visual standard quality of tomatoes, classifying them as *damaged*, *old*, *ripe*, and *unripe*. The theme of product freshness is also explored in [260], [100], [84]. [260] introduced a DL-based model to classify meat freshness into fresh, half-fresh, and spoiled categories, incorporating Grad-CAM++ to support transparent decision-making. [100] presented an InceptionV3 model combined with LIME for efficient and transparent classification of chicken meat freshness, which, when integrated with a robotic arm, enhances automation and food safety in poultry processing. [84] utilized CNN-based models to predict the quality of seabream—categorized as fresh, moderate, or spoiled—based on eye and gill images taken under refrigerated conditions, incorporating LIME and Grad-CAM for model interpretability.

Unlike the previous study, two studies used XAI techniques to analyze **spectral data** to address sensory characteristic problems. [245] developed a CNN model to classify beef freshness using myoglobin data and reflectance spectra, achieving high F1-scores. Grad-CAM highlighted key wavelength regions, confirming myoglobin's importance in freshness classification. The method demonstrated robustness against environmental factors, indicating strong industrial potential. Similarly, [222] used surface-enhanced Raman spectroscopy and a CNN-based model, the Dual-Branch Wide Kernel Network, to classify bacterial signals.

3.6.2 Numerical Explanation

Three works address Sensory Characteristics with the goal of explaining model outputs through numeric explanations, despite their differing approaches and applications based on **tabular data**. [154] focused on predicting boar taint, an undesirable taste and odor found in the meat of male pigs. Using CATBoost, a tree-based ensemble model, the authors achieved peak performance. SHAP analysis identified key factors correlated with boar taint, including feed type, ventilation system, pharmaceutical treatment, and lairage waiting time.[68] developed DL models to classify sweet, bitter, and umami molecules, employing a DNN with molecular descriptors and a graph NN, achieving similar accuracies. SHAP analysis was applied to interpret DNN predictions, revealing key molecular binding properties. [6] developed an ML-based method using various regression models, including XGBoost and RF, alongside FI analysis, to predict aroma partitioning in dairy matrices and support food reformulation efforts.

3.7 Explaining Sustainability and Healthiness

A balanced use of data types and explanation methods is observed in sustainability and healthiness studies, with equal representation of pictorial and tabular data, along with one study utilizing time series data. Table 3.5 summarizes the studies surveyed in the present section.

Table 3.5: Summary of the works introducing applying XAI for sustainability and healthiness surveyed in Section 3.7, according to their data type and explanation type (labelled as "Expl. type").

Works	Data type	Expl. type
[37, 76, 96, 160, 172, 225]	Pictorial	Visual
[75, 112, 119, 120, 196, 252, 265, 300]	Tabular	Numerical
[262]	Time series	Numerical
[230]	Tabular	Rule-based
[245]	Tabular	Mixed

3.7.1 Visual Explanation

The works in this section used Grad-CAM as an XAI technique, confirming its widespread application in explaining solutions to sustainability and healthiness problems using **pictorial data**. [160] used a CNN combined with a feature-based cascade classifier to achieve 83% accuracy in pig face recognition. They employed Grad-CAM to verify that the model focuses on key facial features, offering a cost-effective alternative for animal identification in intensive farming. The paper contributes to the field of animal identification, improving welfare and non-invasive animal management practices. [37] utilized a CNN model, AlexNet, with UAV-based RGB imagery to predict forage biomass, achieving a Mean Absolute Error of 12.98%, with Grad-CAM confirming that the model accurately identified relevant regions for biomass prediction. [96] proposed a CNN model to detect rice phenology stages using smartphone images, reaching 91.30% accuracy. Grad-CAM showed that the model effectively recognized developmental stages, demonstrating the potential of using low-cost tools for real-time agricultural monitoring. [172] introduced MSANet, a model combining multiscale attention and CNN layers for fruit recognition. Grad-CAM was used to interpret the model's decisions, ensuring effective feature identification for robust fruit classification across applications. This work advances waste reduction through automated fruit detection, promoting environmental sustainability. Similarly, [76] applied a ViT model for plant seedling classification and used attention heatmaps to provide insights into the model's decision-making process. Lastly, [225] developed a CNN-based system as an automated method for evaluating the precision of agricultural sprayers by detecting spray deposits, eliminating the need for manual tracers or water-sensitive papers. The study also employed an XAI pipeline—specifically Grad-CAM and Grad-CAM++—to interpret the CNN's decision-making process, revealing key spatial filtering methods used for classification.

3.7.2 Numerical Explanations

The studies explored the application of ML and XAI techniques in health, food, and agriculture using **tabular data**. [75] employed an RF model, using SHAP values to assess the impact of phenol-enriched olive oils on cardiometabolic health in hypercholesterolemic individuals. The study found that phenol-enriched oils sig-

nificantly reduced serum metabolites associated with cardiovascular risk, indicating their potential as a treatment for cardiometabolic diseases. [300] predicted Oral Food Challenge (OFC) outcomes for diagnosing food allergies, with RF and Learning Using Concave and Convex Kernels models achieving high accuracy in identifying egg, peanut, and milk allergies. SHAP analysis highlighted key clinical factors, such as *Immunoglobulin E* levels, as important predictors of OFC outcomes. [120] combined genomic and environmental data to predict wheat yield using advanced DL frameworks.

DeepLift [248] analysis revealed that environmental factors were more influential than genetics, highlighting the importance of integrating both data types for crop variety development. [119] integrated ML and DL models—including SVM, RF, and NNs—with LIME and SHAP to provide a transparent and efficient solution for crop yield prediction, focusing on automating agricultural processes and promoting sustainability. [112] exploited ML-based techniques integrated with LIME and SHAP to predict cattle behavior using sensor data collected from eighteen cows via accelerometers and pressure sensors, classifying behaviors into *Other behavior*, *Ruminating*, and *Drinking/Eating*. [196] applied an RF model to predict almond shelling fraction using genotype data, with SHAP analysis offering insights into the genetic markers influencing shelling fraction, thereby supporting informed breeding strategies. [265] proposed the use of a sensing agricultural robot that collects data such as temperature, humidity, and UV index to automatically forecast mulberry plant diseases by monitoring environmental conditions over time, leveraging LightGBM for prediction and SHAP for interpretability. Finally, [252] introduced a real-time irrigation management system for paddy fields, utilizing a hybrid and ensemble feature extraction approach combined with a Federated Learning-based framework, enabling decentralized learning for localized decision-making while preserving data privacy; SHAP was employed to enhance model interpretability.

Considering **time series**, [262] proposed several ML models to predict individual pig growth trajectories from group-level weight data, reducing reliance on traditional, costly Radio Frequency Identification tracking. The RF model performed best, with an average Root Mean Square Error of 2.26 kg per pig. SHAP analysis highlighted weight and time differences as key predictors, supporting ML as a cost-effective alternative for growth estimation.

3.7.3 Rule-based Explanation

Tabular data was explored in [230], where the authors proposed a system utilizing IoT data, encompassing crop types, soil characteristics, and weather conditions—to monitor the agricultural environment and alert farmers about necessary actions to maintain optimal crop conditions. This method, based on fuzzy logic and integrated with ML algorithms, detects anomalous data resulting from security breaches or hardware malfunctions.

Results indicated that the system effectively increased crop yields through real-time monitoring and decision-making based on IoT insights. The fuzzy logic framework enhanced system interpretability, making it user-friendly for farmers. Tested on maize, the system demonstrated high interpretability, accurate anomaly detec-

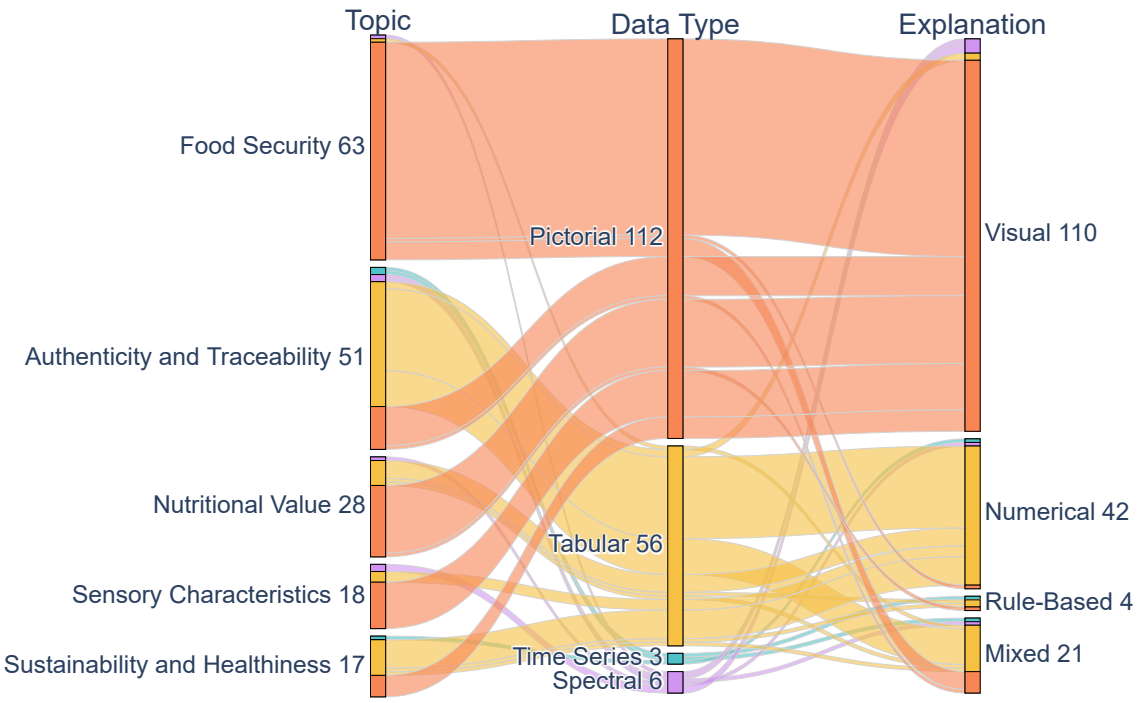


Figure 3.3: Distribution of works surveyed. While the distribution across topics is rather uniform, most of the works we survey concentrate on pictorial data and visual explanations, while a smaller portion of research deals with tabular data and numerical explanations.

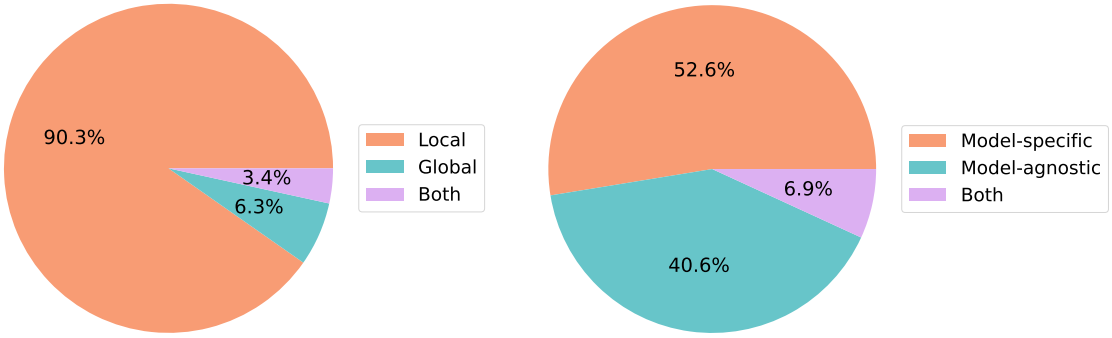
tion, and reliability in triggering appropriate actions.

3.7.4 Mixed Explanation

In the field of Sustainability and Healthiness, only one study applies XAI techniques in conjunction with **tabular data**. [209] introduced AgriUXE, a digital platform that integrates XAI with multimodal data to enhance decision-making in smart farming, bridging the gap between AI-based agricultural solutions and farmers' understanding by providing tailored explanations based on IoT sensor data, remote sensing, and predictive models. The authors presented an effective case study in viticulture by integrating various AI-based methods with multiple XAI techniques, including LIME and SHAP.

3.8 Conclusion

In compiling this literature, we observed that only a small fraction of studies that employ AI also integrate XAI methods. This can be attributed to researchers' focus on developing highly efficient and accurate models to solve the proposed problems. Current food research aims to identify new applications and refine existing models to enhance accuracy. The need for model interpretability becomes less urgent once



(a) (b)

Figure 3.4: Distribution of surveyed papers that utilize global XAI techniques versus local XAI techniques (a), and the percentages of papers that apply model-agnostic XAI techniques compared to model-specific XAI techniques (b). In both charts, the *Both* sector indicates that there are works employing multiple techniques of different types. Observing the first graph, it becomes evident that most of the studies reviewed employ local XAI) techniques, such as the widely recognized SHAP and Grad-CAM. In contrast, global techniques appear to be less commonly used. The second graph confirms this observation, since SHAP is model-agnostic, while Grad-CAM is model-specific, highlighting an even distribution of the two most popular techniques across the various works.

satisfactory performance levels are achieved.

The papers expose that pictorial data are the most frequently used data type, followed closely by tabular data, which is also widely utilized, as shown in Figure 3.3. It is important to note that time series and spectral data, which are widely used in physicochemical analysis, have not been extensively explored with XAI techniques. The prevalence of pictorial data can be attributed to several factors, which can be detected by observing the second half of the plot. XAI techniques that provide visual explanations, such as CAM-based methods, are widely employed in the literature, as noted by [274]. These techniques are highly popular because they provide readily interpretable visual explanations, often as heatmaps, making them ideal for users with limited experience in the analyzed data who still need an intuitive, immediate understanding of the decision-making processes of the image analysis model. This utility justifies why a significant portion of the surveyed papers rely on them, consequently requiring pictorial data. In contrast, techniques that offer numerical explanations, though popular, are not as easily interpretable and are therefore primarily used for analyzing tabular data. Rule-based explanation techniques, on the other hand, are less common and thus less frequently exploited.

Figure 3.4a highlights a clear preference for local methods over global ones. This preference is driven by the popularity of techniques like LIME, SHAP, and Grad-CAM in the reviewed works, all of which are local methods. These methods are simple to apply, offer easily interpretable explanations, and are particularly useful for understanding the model's decision-making in individual cases. In contrast, global methods, which are more suited for gaining an overall view of the model's

decision-making process, are less frequently used due to their complexity, especially when applied to highly intricate models.

The prominence of model-specific techniques, as shown in the Figure 3.4b, is largely due to the widespread use of CAM-based methods. The analysis of the papers suggests that CNNs are the most commonly used approach for pictorial data, while CAM-based techniques are the most straightforward choice for explaining these models. In contrast, LIME and SHAP, the other two most commonly used methods, are model-agnostic. In this case, there is no clear preference between the two types of XAI; instead, the focus tends to be on certain specific techniques.

Lastly, it is worth noting that few papers employ more than one XAI technique, which limits the understanding of the model's decision-making process to a partial view. This is especially true when using local techniques, which provide explanations for individual samples without offering insight into the model's broader decision-making patterns.

Chapter 4

Decision Predicate Graphs: Enhancing Interpretability in Tree Ensembles

In this Chapter, we introduce the DPG, a novel model-specific XAI technique that provides a global explanation of tree-based ensemble models.

The literature review of the previous Chapter shows that, within food quality analysis, applications that integrate explainability are limited and rarely exploit tabular data, despite its value for representing the physicochemical properties of food products. Tabular data also enables the use of tree-based ensembles, which are efficient and high-performing. Moreover, among current explainability methods, approaches that deliver global explanations are comparatively scarce. Global explanations are preferable in contexts like food quality analysis because they establish stable criteria at the model level for evaluating effectiveness across different batches and product lines. Having an overview of the model's entire decision-making process is essential for identifying errors in model learning, recognizing potential biases, or confirming the reliability of results, regardless of specific cases. In contrast, local explanations, while useful for individual predictions, may be inconsistent across similar samples, making them less reliable for guiding interventions at the process level. For these reasons, the technique proposed in this Chapter advances the state of the art in XAI.

Section 4.2 surveys existing methods that provide global explanations for tree-based ensembles. In Section 4.3, we formalize DPG, detailing its construction and the insights available from the resulting explanation. Section 4.4 demonstrates the method on two case studies to illustrate its functioning, benefits, limitations, and possible improvements.

The content of this Chapter has been published in [10]

4.1 Introduction

Artificial intelligence, although still under strong development, is now a consolidated and widely used tool. This is thanks to the continuous growth of computing power, which allows the use of increasingly complex and computationally expensive

ML methods. The challenges presented by modern-world problems are growing in complexity as well as the proposed solutions.

Dealing with large quantities of data and frequently encountering unbalanced datasets are still significant obstacles in addressing many real-world issues; however, tree-based ensemble algorithms offer several advantages in overcoming these challenges, including robustness to noise and outliers, scalability to large datasets, automatic handling of missing values, and the ability to capture complex relationships and interactions within the data [158, 171].

As described in Section 2.2, the process of learning tree-based ensemble models involves training multiple decision trees and aggregating their predictions to enhance performance and generalization across a wide range of tasks. This advantage comes with a corresponding increase in the difficulty of interpreting the decision process, which grows with the number of trees (see Section 2.3). This well known black box property poses a challenge for developers and users alike. In this setting, we introduce the DPG as a method for providing a global explanation of ensemble behavior.

Drawing inspiration from the expanding theme of XAI, we designed a graph structure to tackle transparency and explainability challenges inherent in tree-based ensemble models. This facilitates a better understanding of the intricate choices underlying these ML models. DPG is created with inspiration from the concept of aggregating RF [31, 106], as introduced by Gossen and Steffen [89]. The approach proposed in [89] suggests visualising the decisions within the RF by combining the branches of the tree base learners into a single and compact decision diagram. The concept behind DPG is to convert a generic tree-based ensemble model for classification into a graph, a defined and studied structure with known properties. In this graph, nodes represent predicates, i.e., the feature-value associations present in each node of every tree, while edges denote the frequency with which these predicates are satisfied during the model training phase by the samples of the dataset. The DPG structure enables comprehending the choices made by the model, enhancing transparency and understandability. Moreover, it allows the exploitation of graph properties to develop metrics and algorithms facilitating the analysis of the ensemble model. This, in turn, aids in understanding the decisions it makes, easing the task of visualising the graph which can be vast and complicated for larger models with numerous tree base learners.

DPG serves as a model-specific tool offering a comprehensive interpretation of tree-based ensemble models. It provides descriptive metrics that enhance the understanding of the decisions inherent in the model, offering valuable insights. This tool proves particularly useful for models that are *a priori* considered satisfactory in terms of performance.

Our work contributes in the following ways:

- we introduce DPG, a novel interpretability structure that transforms an opaque-box tree-based ensemble model into an enriched graph;

- we present the algorithm used to create DPG, accompanied by pseudo-code to enhance understanding and facilitate replication, complete with its asymptotic complexity;

- we provide the interpretation of three metrics from graph theory, enriching the model comprehension and gaining insights;

- we demonstrate the use of the proposed method through two case studies: the application of DPG to two RF models, respectively, on the Iris dataset [79] and a challenging dataset.

It is important to highlight that these results are achieved in a generic fashion, utilising a standard classifier on well-established datasets. Significantly, we intentionally avoided incorporating scenario-specific heuristics. Therefore, we posit that our aggregation approach has the potential for widespread application across a diverse range of related scenarios.

4.2 Literature Review

As we introduced in Section 2.3, XAI tools, especially those providing global interpretations, become valuable instruments for understanding tree-based ensemble models. These models are widely used in addressing diverse problem domains, as highlighted in several surveys [8, 49, 93]. As a result, the number of studies exploring model-specific techniques designed for ensembles of trees has also increased. Below, we outline the key contributions that represent the current state of the art in which our technique is situated. We analyze the main differences to highlight the novelty of our approach compared to existing ones, noting that an analytical comparison can be challenging due to the diversity of explanations offered in the literature.

The first significant study is proposed by Mashayekhi and Gras [163]. The authors introduced *RF+HC*, an approach that employs a hill climbing algorithm in RF to search for a decision set. This rule set reduces the number of decisions dramatically, which significantly improves the comprehensibility of the underlying model built by RF.

Similarly, Hara and Hayashi [99] exploit Bayesian model selection to extract the decision set. These approaches share similarities with our method; however, our proposal extends beyond the extraction of decisions from the RF. Visualising the decisions of tree-based ensemble models and simultaneously complementing them with metrics developed by graph theory makes DPG more adaptable and holistic. This approach allows for obtaining insights into the model beyond just decision information.

Zhao et al. [304] proposed *iforest*[1], a visual analysis system specifically designed to interpret RF models and their predictions. They built a feature view to illustrate the relationships between input features and outcome predictions and proposed a design that summarises multiple decision paths based on feature occurrences and ranges, allowing users to explore and understand the partitioning logic of these paths. The *iforest*, like numerous visualisation systems, faces significant challenges related to scalability and interpretability when dealing with large ensembles. To

[1]In the referenced work, the author refers to this technique as "iforest," and we will use that term only in this section. It should not be confused with Isolation Forest, which is also commonly shortened to "iForest" and is used throughout the thesis.

overcome this challenge, our approach does not solely rely on visualizing the model as a graph, but provides metrics within the explanation that enhance understanding of the decision-making process.

Hatwell et al. [102] contributed with *Collection of High Importance Random Path Snippets* (CHIRPS), a method that incorporates the explanation of RF classification for each data instance and extracts a decision path from each tree in the forest, resulting in a set of decisions that elucidate the classification process. However, this method is limited to rules extraction and lacks insight into the model's structure. Additionally, it does not incorporate metrics to explain the logic of the tree-based ensemble model.

Another technique focused on visualising decisions underlying RF models is introduced by Neto and Paulovich [190]. *Explainable Matrix* (ExMatrix) employs a matrix-like visual metaphor, where rows represent decisions, columns denote features, and cells encapsulate decision predicates, thereby facilitating the scrutiny of models and the audit of classification outcomes. The visualisation capability of ExMatrix for global visualisation has limitations in terms of scalability because the number of decisions increases significantly with the number of trees in large ensembles. Moreover, ExMatrix layouts can rapidly become challenging to explore, while the complexity of the model increases. As mentioned earlier, our approach enables us to retrieve information without relying solely on visualisation.

Dedja et al. [62] introduced an approach denoted as BELLATREX, which is designed to explain the forest predictions for a given test instance by a set of logical rules based on the features of the dataset. However, a potential limitation lies in the computational complexity of the approach. While explaining a single prediction is quick, applying the method to a complete dataset becomes computationally expensive. Furthermore, BELLATREX focuses on instance-level explanations rather than providing a global perspective. In addition, BELLATREX uses clustering techniques to simplify the representation and decision logic, whereas our approach uses graphs to avoid simplifications that can lead to loss of information.

Various studies [63, 91, 270, 310] proposed several tree similarity metrics through the process of clustering based on tree representations. However, these approaches, while beneficial for interpretability, require simplifying the model through techniques such as selection, pruning, and frequency analysis, which can result in information loss.

A number of works [89, 182, 186, 250] established a connection between tree-based ensemble models and graph theory. The foundational concept of these techniques has been explored by [80, 113, 188, 201, 259, 278, 312]. These works established the theory that introduces the transformation of decision trees into graphs, aiming for more efficient and non-redundant tree structures. Nakahara et al. [186] and Silva et al. [250] works focus on performance optimisation, with Gossen and Steffen [89] and Murtovi et al. [182] being the sole contributors that employed these techniques for interpretability purposes.

In particular, Gossen and Steffen [89] introduced the Algebraic Decision Diagram (ADD), aiming to transform tree-based ensemble models into bipartite graphs. The ADD serves as an alternative construction to RF, providing an additional predictive model that functions as a surrogate. ADD proves valuable for specific aspects of

interpretability, such as outcome explanation, logic of *majority vote*, and visualising the path. Acting as a surrogate model, their primary contributions lie in providing a simple, optimised model. In comparison, our objective is to extend their ideas by utilizing metrics derived from graph theory. This goes beyond the transposition of tree-based ensemble models into a graph; rather, it involves using graph theory and its associated tools to gain insights into the functionality of the models.

ForestGUMP, an online tool developed by Murtovi et al. [182], is designed for generating ADDs. This tool provides valuable information such as graph visualisation, hypothetical sample path display, and logic of majority vote. However, it has some limitations in terms of the number of usable trees (only 20) and in visualising problems that involve multiple features and classification choices, making graph navigation complex. In our work, while we enable visualisation, our focus is on utilising graph-related metrics, and we do not incorporate the simplification of the analysed models.

4.3 Decision Predicate Graphs

DPG converts complex ensemble models into a graph structure, where nodes represent predicates made by the model and edges denote the occurrence of these predicates during model training.

In this section, we present the formalisation of DPG, elucidate the algorithm employed in its development through pseudo-code and its asymptotic complexity, and provide an exposition of metrics and properties for comprehending the tree-based ensemble model. Moreover, we outline the advantages of metrics and articulate why they can serve as a complementary aid to graph visualisation, particularly in overcoming its inherent limitations.

As previously mentioned in Section 4.1, DPG is tailored for tree-based ensemble models designed specifically for classification tasks.

4.3.1 Definition

Let $\mathcal{M}_n$ be the tree-based ensemble model consisting of n tree base learners $T(x;\Theta_b)$, where x is a generic sample, and Θ_b characterizes the bth learner in terms of split variables, cutpoints at each node, and terminal-node values. More specifically, Θ_b includes:

- all the splitting conditions associated with each jth internal node n_{bj} based on a specific feature f_{bj} and a threshold (for numerical features) or a set of possible values (for categorical features) t_{bj};
- the values assigned to leaf nodes c_b.

Let $\mathcal{D}$ be the training set on which $\mathcal{M}_n$ is trained. Every base learner is trained on a dataset $\mathcal{D}_b$, where $\mathcal{D}_b$ is a subset of $\mathcal{D}$. We define $\mathcal{O}$ as the set of logical operations: $\mathcal{O} = \{\leq, >, =, \neq\}$.

The predicate set $\mathcal{P}(\mathcal{M}_n)$, for an ensemble method $\mathcal{M}_n$ is the set obtained by the union of the set of all the triples $p = (f_{bj}, o, t_{bj})$, where $o \in \mathcal{O}$, and f_{bj}, $t_{bj} \in \Theta_b$,

and the set of all leaf nodes c_b, for all n tree base learners of $\mathcal{M}_n$. The triples p are called decisions, while the elements of $\mathcal{P}(\mathcal{M}_n)$ are called predicates.

A DPG (DPG($\mathcal{M}_n$)) for a model $\mathcal{M}_n$ is a directed weighted graph $(\mathcal{P}, E)$ where:

- $\mathcal{P}$ is the set of nodes, which corresponds to the predicate set $\mathcal{P}(\mathcal{M}_n)$;
- E is the set of edges, where each directed edge connects two predicates if and only if there exists an element in the training set $\mathcal{D}$ that satisfies both conditions specified by the predicates in an immediately consecutive manner. The weight of the edge corresponds to the number of training set elements satisfying these consecutive conditions.

For conciseness, from this point onward, we will use the acronym DPG to refer to the graph, indicating that it is constructed based on a model.

4.3.2 From Ensemble to a DPG

We introduce an algorithm, outlined in Algorithm 1, for constructing the DPG based on a tree-based ensemble model, by traversing all tree base learners with the training samples.

Algorithm 1: Construct DPG from Ensemble Tree Model

Input: Ensemble tree model $\mathcal{M}_n$, Training set $\mathcal{D}$
Output: DPG($\mathcal{M}_n$)

```
Initialise empty set DPG(M_n);
foreach T (learner) in M_n do
    Initialise empty predicate set P and edge set E;
    foreach x (sample) in D do
        Initialise empty predicate set P_x and edge set E_x;
        // To obtain the predicates path for x on the tree T
        (P_x, E_x) ← TRAVERSING(T, x);
        Add (P_x, E_x) to P and E;
DPG(M_n) ← AGGREGATING(P, E);
return DPG(M_n);
```

The algorithm iterates over each base learner in the ensemble tree model $\mathcal{M}_n$ and each training sample x in the training set $\mathcal{D}$.

To clarify, the `TRAVERSING` function follows the predicate path of a particular input sample x through the decision tree T, starting from the root node and navigating to the appropriate leaf node based on the feature values of x. Meanwhile, the `AGGREGATING` function processes the predicates and edges obtained from `TRAVERSING` the decision trees for all samples into a single graph representation, DPG($\mathcal{M}_n$), by taking the union of $\mathcal{P}$ and computing the frequency of elements of E.

The algorithm presents a systematic methodology for constructing DPG. The overall asymptotic complexity can be formally expressed as follows:

$$O(b \times s \times (k + k^2)) = O(b \times s \times k^2)$$

where:

- b is the number of learners in the ensemble,
- s is the number of samples in the training set, $|\mathcal{D}|$, and
- k represents the size of the $(\mathcal{P}_x, E_x)$ processed by the `TRAVERSING` and `AGGREGATING` functions.

This analysis takes into account the linear time complexity $O(k)$ for the `TRAVERSING` function and the quadratic time complexity $O(k^2)$ for the `AGGREGATING` function. Our *Python* 3.10 implementation is accessible here[2].

4.3.3 DPG interpretability

In this section, we enumerate and elucidate some of the advantages that DPG can offer. The graph-based nature of DPG provides significant enhancements in the direction of a complete mapping of the ensemble structure. Weighted directed graphs, such as DPGs, are studied structures with well-established properties that enable the identification or construction of useful metrics and algorithms. It is crucial to emphasise that all the observations presented in this section are valuable for comprehending and analysing the obtained model.

Visualisation.

DPG provides an immediate advantage by allowing the visualisation of the entire tree-based ensemble models through a single comprehensive graph. Similar to the idea proposed by Gossen and Steffen [89], consolidating all individual basic learners within the model into a unified graph provides a holistic representation of the decision-making process. This visualisation not only elucidates the decisions made by the learners but also reveals the intricate relationships between them. Consequently, it facilitates a comprehensive understanding of the utilised features and, more importantly, the associations between features and their values that enable the model to accurately classify a sample into a specific class.

Another noteworthy aspect of DPG lies in the representation of weights for pairwise nodes. This feature enables a discerning analysis of the most significant path through predicates, shedding light on decisions consistently employed by numerous learners or across multiple samples. This insight not only highlights the prevalence of certain decisions but also opens avenues for targeted enhancement strategies, focusing on those influential aspects within the model.

Moreover, by traversing all the possible paths between predicates in reverse, starting from one of the classes, we can discern the essential characteristics that a sample must possess to be classified into a specific class. This capability facilitates the *a priori* elimination of certain elements from the dataset when considering the particular class.

Nevertheless, we acknowledge that while visualisation is a valuable tool, its effectiveness diminishes with an increasing number of tree base learners. A multitude of tree base learners implies an increase in decisions and, consequently, an abundance

[2] https://github.com/LeonardoArrighi/DPG

of predicates. As a result, the size of the graph, in terms of nodes, grows proportionally with the model's scale. This expansion can render the graph illegible or impractical to visualise due to its intricate complexity. To address this challenge, we provide additional tools that complement the visualisation, aiding in the extraction of model properties and facilitating a more comprehensible understanding.

One approach is based on the desire and feasibility of determining the specific characteristics a sample must exhibit to be assigned to a particular class. Taking inspiration from the *outcome explanation problem* introduced by Gossen and Steffen [89], to enhance the immediacy and effectiveness of this analysis, we provide an aggregation of predicates, referred to as *constraint*, which represent intervals associated with the features of each class. The constraints are defined as follows: for a given class identified in the DPG, we list all nodes connected by a path originating from the node itself and culminating in the class. For each feature within the node predicates, we delineate the most extensive possible interval using the values associated with the features. This interval is defined by two endpoints. The minor endpoint is the smallest value within the set of values less than the feature, while the major endpoint is the largest value within the set of values greater than the feature. If either of these two sets is empty, the interval is deemed infinite. Each class has its constraints for every feature contributing to the classification of the samples.

It is important to note that constraints are not a substitute for visualisation; instead, they offer insights resembling those promoted by it. It is anticipated that additional graph measures could complement insights from constraints, providing similar interpretations as visualisation.

Centrality.

The centrality of a node is defined as a number or rank corresponding to the node position within the network. By observing centrality, we can make considerations that allow us to better understand the process hidden in the ensemble method. The notion of centrality encompasses a wide range of metrics. In this section, we explore those metrics that offer the most insightful information.

According to Brandes [30], we define the Betweenness Centrality (BC) of a node as the fraction of all the shortest paths between every pair of nodes of the graph passing through the considered node. Let DPG $= (\mathcal{P}, E)$ be the graph and $s, t, v \in \mathcal{P}$ three vertexes of DPG, we can denote with $\sigma(s, t)$ the number of shortest paths between s and t and with $\sigma(s, t|v)$ the number of shortest paths between s and t passing through v. Then, the BC of the node $v \in \mathcal{P}$ is defined as:

$$BC(v) = \sum_{s,t \in \mathcal{P}} \frac{\sigma(s, t|v)}{\sigma(s, t)}.$$

All details and observations about BC can be read in [30]. BC serves as a relevant metric for gaining a deeper insight into the significance of decisions within the ensemble model. We can observe that a node with a higher BC value has a more significant influence on the flow of information within the graph; nodes with high BC can be considered potential bottleneck nodes because they play a crucial role in facilitating interactions between different parts of the DPG. For this reason, we

can assert that these nodes are meaningful to understanding the tree-based ensemble models: in all tree base learners, the decision contained in the node is essential to classify the elements of the dataset. We highlight that the significance extends beyond the characteristic itself; it encompasses the value associated with it.

According to Mones et al. [177], we define the Local Reaching Centrality (LRC) of a node v of the DPG as the proportion of other nodes reachable from node v via outgoing edges. LRC can be generalised to weighted graphs by measuring the average weight of a given directed path starting from node v (more details are available on [177]). The LRC serves as a metric for assessing the importance of DPG's nodes. It gauges the extent to which decisions contained in these nodes are employed by diverse tree base learners for classifying samples in the training set. This, in turn, reflects the importance of these decisions in the classification of new samples. The LRC offers a comprehensive perspective on the concept of FI by extending its definition to encompass the values associated with features across various decisions. Additionally, the prominence of paths between highlighted predicates indicates their frequent utilisation, providing insights into how new samples can be classified with fewer decisions.

Community.

While there is no single definition of a community, we can observe structures similar to communities in the DPG. According to Radicchi et al. [216], we define a *community* as a subset of nodes of the DPG characterised by dense interconnections between its elements and sparse connections with the other nodes of the DPG that do not belong to the community. Based on the properties of DPG, we employed *asynchronous label propagation* algorithm, proposed by Raghavan et al. [217], to detect communities within the graph. The core concept of the algorithm involves each graph node determining its community membership based on the majority of its neighbours. The algorithm comprises a series of steps: each node initially possesses a unique label. As these labels diffuse through the graph, closely connected groups of nodes converge on a common label. These consensus groups then expand outward until further expansion becomes impractical. After this label propagation process, nodes sharing the same labels are identified as belonging to the same community. This process is iterated until each node in the network aligns its label with the community that includes the maximum number of its neighbouring nodes. The algorithm is defined asynchronous, as each node receives updates without waiting for updates on the remaining nodes. Identifying communities in the DPG provides insights into the ensemble model: visualising them allows us to discover groups of nodes that similarly contribute to the classification of samples.

By employing the asynchronous label propagation algorithm, we observe that each formed community is associated with a class. Within these communities, the features utilised by the ensemble model to classify samples belonging to the community's class are emphasised. Once again, the association between features and values plays a key role, highlighting the specific decisions made by the learners. To quote Raghavan et al. [217]:

> Communities in social networks can provide insights about common char-

acteristics or beliefs among people that make them different from other communities.

Similarly, we observe that communities within the DPG offer a valuable understanding of the characteristics for samples to be assigned to a particular community class. This intuition extends to identifying predominant features and those that play a marginal role in the classification process. Moreover, it is noteworthy that communities also provide insights into the entire dataset and the complexity of the problem. A community comprising a substantial number of nodes, each associated with different predicates often involving distinct features, indicates that the model makes diverse decisions to assign samples to the community class. This implies that the model encounters challenges in classifying samples for this particular class, and data from different classes are not easily distinguishable within the dataset.

Finally, communities, functioning as sub-graphs, can be used to visualise the decisions made in the ensemble model, enabling the identification of a specific class. This replaces the complex illustration of the DPG, especially when we are focused on visualising a single class and dealing with many tree base learners. We summarised the utility of discussed properties and metrics in Table 4.1.

4.4 Empirical Results and Discussion

In this section, we demonstrate the effectiveness of DPG to the well-known Iris dataset [79] and a synthetic multiclass dataset[3]. Each experiment in this section was conducted using the RF classifier, with variations limited to the number of tree base learners. The implementation is available here[4]. Finally, we discuss potential enhancements to DPG and explore further development opportunities.

4.4.1 DPG: Iris insights

The first case study concerns the classification of the Iris dataset [79]. The simplicity, manageability, versatility, and relevance of this dataset make it an interesting and relevant resource for discussions and demonstrations of interpretability in ML. The dataset comprises measurements of sepals and petals for iris flowers encompassing three distinct species with a total of four features and three classes. The RF was selected as the tree-based ensemble model due to its well-established reputation and high-performance capabilities. To conduct the classification, we partitioned the dataset into training and test sets, following a 80-20% proportion, respectively. A seed value of 42 was established for randomness control, and the number of base tree learners was set to 5. The RF performances, evaluated on the test set, are summarised in the confusion matrix in Table 4.2. The model demonstrates 100% accuracy.

After training the model, we applied the algorithm outlined in Section 4.3.2 to obtain the DPG, which can be visualised in Figure 4.1. Then, we can analyse the obtained graph using the metrics and algorithms proposed in Section 4.3.3. It is

[3] https://github.com/LeonardoArrighi/DPG/tree/main/datasets
[4] https://github.com/LeonardoArrighi/DPG

Table 4.1: Summary of DPG properties and metrics, and their utility in offering insights into tree-based ensemble models.

Property	**Definition**	**Utility**
Constraints	The intervals of values for each feature obtained from all predicates connected by a path that culminates in a given class.	Calculate the classification boundary values of each feature associated with each class.
BC	Quantifies the fraction of all the shortest paths between every pair of nodes of the graph passing through the considered node.	Identify potential bottleneck nodes that correspond to crucial decisions.
LRC	Quantifies the proportion of other nodes reachable from the local node through its outgoing edges.	Assess the importance of nodes similarly to FI, but enrich the information by encompassing the values associated with features across all decisions.
Community	A subset of nodes of the DPG which is characterised by dense interconnections between its elements and sparse connections with the other nodes of the DPG that do not belong to the community.	Understanding the characteristics of nodes to be assigned to a particular community class, identifying predominant predicates, and those that play a marginal role in the classification process.

Table 4.2: Confusion matrix depicting the performance evaluation of the RF model with 5 base tree learners on the test set.

	Prediction		
Ground truth	Class 0	Class 1	Class 2
Class 0	19	0	0
Class 1	0	13	0
Class 2	0	0	13

important to note that the DPG leads to the calculation of both global metrics, referring to the overall graph, and metrics at the level of individual nodes.

To illustrate the effectiveness and one of the advantages of employing DPG, we highlight the constraints for the different classes in the Table 4.3. The class-specific constraints delineate the necessary characteristics a sample must exhibit to be assigned to that particular class by the tree-based ensemble model. This insight

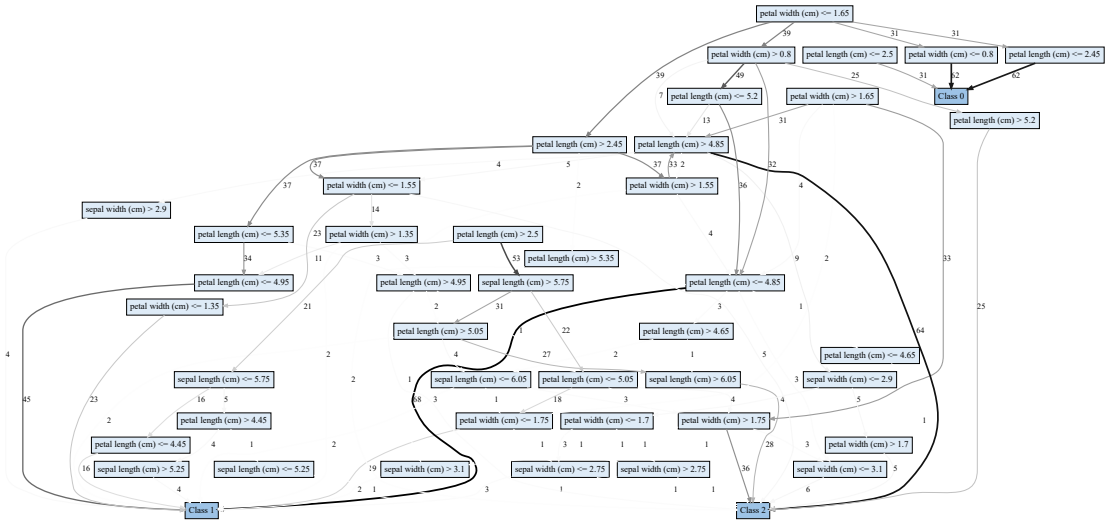


Figure 4.1: DPG of the RF composed of 5 tree base learners trained on Iris dataset.

Table 4.3: Constraints for each class based on the DPG for an RF model within 5 tree base learners.

Class	Constraints
0	`petal width (cm)` ≤ 1.65 `petal length (cm)` ≤ 2.50
1	$5.25 <$ `sepal length (cm)` ≤ 6.05 $0.80 <$ `petal width (cm)` ≤ 1.75 $2.45 <$ `petal length (cm)` ≤ 5.35 $2.75 <$ `sepal width (cm)` ≤ 2.90
2	$5.75 <$ `sepal length (cm)` ≤ 6.05 $0.80 <$ `petal width (cm)` ≤ 1.75 $2.45 <$ `petal length (cm)` ≤ 5.35 $2.75 <$ `sepal width (cm)` ≤ 3.10

contributes to a better understanding of how the model utilises features for effective classification.

The first metric under discussion is the BC of the nodes, as depicted in Table 4.4, where we identify potential bottleneck nodes. These nodes encapsulate significant information, particularly representing decisions made by numerous tree base learn-

Table 4.4: Top eight predicates by evaluating their BC obtained from the DPG based on an RF model consisting of 5 tree base learners.

Predicate	**BC**
`petal length (cm) > 4.85`	0.053
`petal length (cm) <= 4.85`	0.036
`petal width (cm) > 1.55`	0.034
`sepal length (cm) <= 6.05`	0.032
`petal length (cm) > 4.95`	0.028
`petal length (cm) > 4.65`	0.022
`petal width (cm) <= 1.75`	0.022
`petal width (cm) <= 1.55`	0.021

Table 4.5: Comparison of the top eight predicates by evaluating their LRC obtained from the DPG based on an RF model consisting of 5 tree base learners (Table 4.5a), alongside the FI of the same model (Table 4.5b) calculated exploiting MDI algorithm.

(a) LRC evaluation

Predicate	**LRC**
`petal width (cm) <= 1.65`	1.531
`petal length (cm) > 2.45`	0.919
`petal width (cm) > 0.80`	0.874
`petal length (cm) > 2.50`	0.699
`petal width (cm) > 1.65`	0.618
`petal length (cm) <= 5.20`	0.565
`petal width (cm) > 1.55`	0.540
`sepal length (cm) > 5.75`	0.332

(b) FI evaluation

Feature	**FI**
`petal length (cm)`	0.550
`petal width (cm)`	0.373
`sepal length (cm)`	0.054
`sepal width (cm)`	0.023

ers. We quickly discern that the decision associated with `petal length (cm)` and the value 4.85 is pivotal, as it is frequently relied upon by multiple basic learners and is essential for successful classification.

Furthermore, the LRC metric provides additional information. Examining Table 4.5a, we can emphasise the most crucial predicates influencing the decision-making process of the tree-based ensemble model. This includes not only identifying the most frequently used features but also recognising the associated values that lead to significant and divisive splits in the dataset across various basic learners. As observed in Table 4.5, a comparison between the LRC of the nodes and the FI, calculated on the same model on which DPG is based, suggests that the metric may rank the predicates similarly. FI is calculated using the Mean Decrease in Impurity (MDI) algorithm introduced by Breiman [31]. This comparison also provides additional information about the values used in the decisions and the frequency of paths extending the concept of FI.

By employing the global metric community, we identified the presence of three distinct communities. Table 4.6 illustrates the association between each community

Table 4.6: Communities obtained from the DPG based on an RF model composed of 5 tree base learners. The table shows the number of predicates belonging to each community, the number of features in the community nodes, and the class involved in each community.

Community	**# Predicates**	**# Features**	**Class**
Community 1	23	4	1
Community 2	18	4	2
Community 3	4	2	0

and a distinct class obtained by applying the asynchronous label propagation algorithm to the DPG. We can affirm that each node within a community contains decisions that significantly contribute to the accurate classification of samples belonging to a specific class. For instance, when applying the predicates of Community 3 (comprising two features and two predicates) to the test set and traversing from the root node, it achieved 100% accuracy for Class 0, the class delineated in the mentioned community.

Finally, upon comparing the Table 4.6 and the Figure 4.2, we can state that the communities facilitate the comprehension of how the model addresses the classification problem. Examining the number of decisions and features utilised in each community reveals that differentiating between Class 1 (Community 1) and Class 2 (Community 2) poses a greater challenge for the model. This indicates the difficulty in effectively separating samples within the dataset into their respective classes. This difficulty becomes apparent when visualising the dataset across the features, as in Figure 4.2. Conversely, the community encompassing Class 0 (Community 3) consists of fewer predicates, signifying that it is more distinguishable from other classes, as confirmed in the Figure 4.2.

4.4.2 Comparing to the Graph-based Solutions

As outlined in the Section 4.2, Gossen and Steffen [89] and Murtovi et al. [182] conducted studies on the interpretability of tree-based ensemble models exploiting graph structures. We compared DPG with their proposed method by examining their outcomes and potential insights. Using the same tree-based ensemble model we studied in Section 4.4.1 as input, we generated the ADD displayed in Figure 4.3. The first noticeable distinction from DPG lies in the ADD structure, as it forms a bipartite graph. We can observe that the ADD is generated from the trained model, albeit without fully leveraging the training dataset. Consequently, the evaluation of connections between nodes and the assessment of the significance of decisions made by different tree base learners are not fully exploited. This implies that each branch carries equal weight and impact in the diagram, and a classification error significantly influences the structure by introducing an incorrect path. Another crucial difference is that ADD does not provide graph metrics, leaving the user the interpretation of the diagram and potentially missing out on relevant information. Moreover, an additional limitation, as indicated in the studies by [89, 182], involves

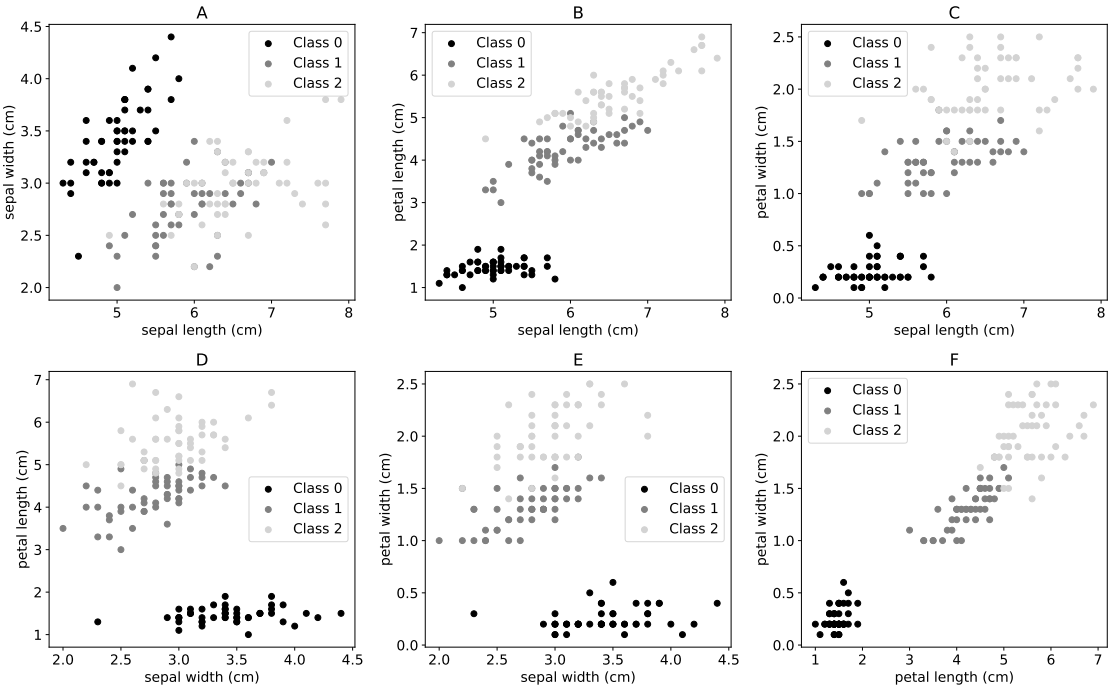


Figure 4.2: Two-dimensional depiction of the Iris dataset, employing feature pairs in each graph for visual representation.

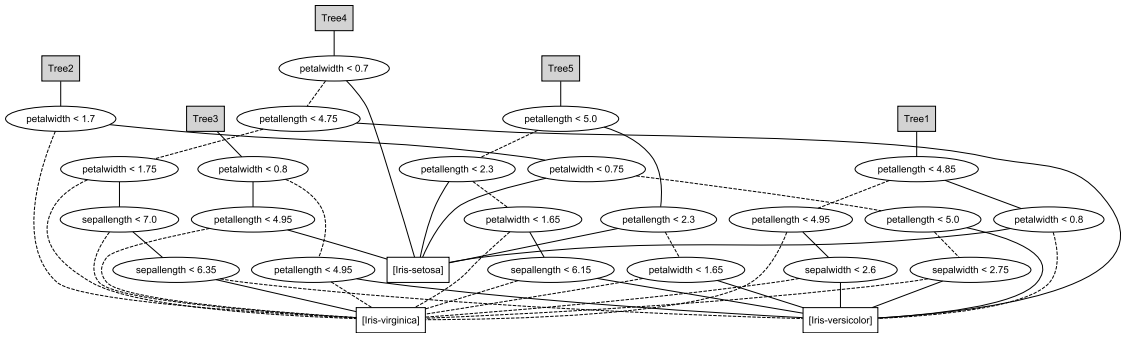


Figure 4.3: ADD of an RF model with 5 tree base learners induced for Iris dataset.

the challenge that emerges when generating ADDs and dealing with large ensembles. Visualisation becomes intricate, even with a modest count of 20 tree base learners. In contrast, DPG allows the computation of both global and local metrics, even with a higher number of tree base learners.

To further examine these aspects, we use two RF models, one with 20 tree learners and the other with 100, to analyse a complex multiclass problem with a dataset comprising 4 classes, 1000 samples, and 16 features. This introduces a four-class problem that was randomly generated. The dataset was created using the `make_classification`[5] function from `scikit-learn`. The following setup has been maintained for training both models. We divided the dataset into training

[5] https://scikit-learn.org/stable/modules/generated/sklearn.datasets.make_classification.ht\gls{ml}

Table 4.7: Confusion matrices of the RF models with 20 tree base learners (RF 20) and with 100 tree base learners (RF 100) tested on the synthetic dataset. To enhance clarity in representation, we abbreviate "Class" as "Cl."

	Prediction (RF 20)				**Prediction (RF 100)**			
Ground truth	Cl. 0	Cl. 1	Cl. 2	Cl. 3	Cl. 0	Cl. 1	Cl. 2	Cl. 3
Cl. 0	38	4	12	6	38	4	14	4
Cl. 1	5	31	3	5	4	32	2	6
Cl. 2	11	2	29	2	5	2	33	4
Cl. 3	10	13	10	19	9	9	5	29

Table 4.8: Constraints for each class based on the DPG for an RF model within 100 tree base learners.

Class 0	Class 1	Class 2	Class 3
$-5.87 < \texttt{F1} \leq 5.74$	$-5.79 < \texttt{F1} \leq 5.72$	$-5.76 < \texttt{F1} \leq 5.72$	$-5.76 < \texttt{F1} \leq 5.72$
$-2.64 < \texttt{F2} \leq 2.63$	$-2.61 < \texttt{F2} \leq 2.61$	$-2.61 < \texttt{F2} \leq 2.61$	$-2.61 < \texttt{F2} \leq 2.63$
$-5.24 < \texttt{F3} \leq 3.75$	$-5.38 < \texttt{F3} \leq 3.75$	$-5.24 < \texttt{F3} \leq 3.75$	$-5.24 < \texttt{F3} \leq 3.75$
$-4.80 < \texttt{F4} \leq 4.37$	$-5.15 < \texttt{F4} \leq 4.37$	$-4.80 < \texttt{F4} \leq 4.37$	$-4.80 < \texttt{F4} \leq 4.37$
$-2.61 < \texttt{F5} \leq 2.44$	$-2.61 < \texttt{F5} \leq 2.71$	$-2.61 < \texttt{F5} \leq 2.44$	$-2.61 < \texttt{F5} \leq 2.44$
$-2.29 < \texttt{F6} \leq 2.58$	$-2.29 < \texttt{F6} \leq 2.58$	$-2.29 < \texttt{F6} \leq 2.58$	$-2.29 < \texttt{F6} \leq 2.58$
$-2.82 < \texttt{F7} \leq 2.47$	$-2.82 < \texttt{F7} \leq 2.47$	$-2.82 < \texttt{F7} \leq 2.47$	$-2.82 < \texttt{F7} \leq 2.86$
$-4.62 < \texttt{F8} \leq 4.74$	$-4.62 < \texttt{F8} \leq 4.74$	$-4.62 < \texttt{F8} \leq 4.74$	$-4.62 < \texttt{F8} \leq 4.74$
$-2.40 < \texttt{F9} \leq 2.59$	$-2.40 < \texttt{F9} \leq 2.59$	$-2.40 < \texttt{F9} \leq 2.71$	$-2.40 < \texttt{F9} \leq 2.59$
$-4.71 < \texttt{F10} \leq 4.32$	$-4.71 < \texttt{F10} \leq 4.32$	$-4.71 < \texttt{F10} \leq 5.43$	$-4.71 < \texttt{F10} \leq 4.32$
$-2.87 < \texttt{F11} \leq 2.77$	$-2.87 < \texttt{F11} \leq 2.77$	$-2.87 < \texttt{F11} \leq 2.86$	$-2.87 < \texttt{F11} \leq 2.77$
$-2.42 < \texttt{F12} \leq 2.37$	$-2.42 < \texttt{F12} \leq 2.37$	$-2.42 < \texttt{F12} \leq 2.37$	$-2.42 < \texttt{F12} \leq 2.37$
$-4.28 < \texttt{F13} \leq 5.01$	$-4.28 < \texttt{F13} \leq 5.01$	$-4.28 < \texttt{F13} \leq 5.01$	$-4.28 < \texttt{F13} \leq 5.01$
$-7.31 < \texttt{F14} \leq 8.33$	$-7.31 < \texttt{F14} \leq 8.33$	$-7.31 < \texttt{F14} \leq 8.33$	$-7.31 < \texttt{F14} \leq 8.33$
$-4.46 < \texttt{F15} \leq 4.19$	$-4.46 < \texttt{F15} \leq 4.19$	$-4.46 < \texttt{F15} \leq 4.19$	$-4.97 < \texttt{F15} \leq 4.5$
$-4.70 < \texttt{F16} \leq 4.21$	$-4.70 < \texttt{F16} \leq 4.21$	$-4.70 < \texttt{F16} \leq 4.52$	$-4.70 < \texttt{F16} \leq 4.21$

and test sets, with an 80-20% ratio. We fixed a seed value of 42 for randomness control to ensure reproducibility. The RF performances, assessed on the test set, are summarised in the confusion matrices presented in Table 4.7. The model with 100 tree base learners shows better performance in every parameter, reaching an overall accuracy of 66%, outperforming the model with 20 tree base learners, which barely reaches 58%.

We emphasise that even in this context, DPG is a useful tool. Both DPG and ADD present intricate visualisations with 20 tree base learners. However, DPG overcomes this obstacle by providing metrics that can still offer valid insights into the model. The first insight is displayed in Table 4.8, where we provide constraints for the four classes of the dataset. Constraints, even in this complex scenario, allow the visualisation of intervals where sample features should be situated for precise classification into their respective classes.

The BC metric helps identifying potential bottleneck nodes. Upon observing

Table 4.9: Top eight predicates by evaluating their BC (Table 4.9a), and top eight predicates by evaluating their LRC (Table 4.9b), both obtained from the DPG based on an RF model consisting of 100 tree base learners.

(a) BC evaluation

Predicate	**BC**
`F15 > 1.17`	0.018
`F15 <= 1.61`	0.015
`F12 > 0.20`	0.015
`F12 > 0.41`	0.014
`F4 <= 0.33`	0.014
`F8 > 0.36`	0.014
`F1 <= -1.10`	0.013
`F11 <= 1.16`	0.013

(b) LRC evaluation

Predicate	**LRC**
`F7 <= 1.62`	15.812
`F1 <= 3.10`	14.475
`F14 > -1.78`	13.313
`F4 > -2.97`	13.158
`F5 > -1.92`	13.065
`F4 > -1.36`	12.989
`F1 <= 2.49`	12.986
`F13 > 1.98`	12.920

Table 4.10: Communities obtained from an RF model composed of 100 tree base learners. The table shows the number of predicates belonging to each community, the number of features in the community nodes, and the class involved in each community.

Community	**# Predicates**	**# Features**	**Class**
Community 1	7767	16	2
Community 2	2149	16	0
Community 3	2351	16	3
Community 3	2100	16	1

Table 4.9a, we can see that there is not a large difference between the BC values associated with the predicates. Therefore, we can assume that there are no bottleneck nodes.

Examining the information provided in Table 4.9b, the LRC underscores which predicates significantly impact the decision-making process of the ensemble model.

Another insight can be obtained by employing the global metric community. In this scenario, we identified the presence of four distinct communities, displayed in Table 4.10. We note that each community contains a distinct class. Furthermore, upon observing the table, we can conclude that each community exhibits a high number of involved features and predicates, confirming the complexity of the classification problem.

4.4.3 Potential Improvements

Several avenues await exploration in the future. The primary aim is to reduce the computational cost of DPG, as many real-world problems involve large datasets that do not scale well with the current implementation of DPG. Expanding the

application scope of DPG is another key goal, including its utility in explaining models relevant to regression-type problems. Given DPG's applicability to any model and dataset, we aim to introduce new tests and use cases to delve deeper into the method. This includes proposing applications to novel datasets and exploring their compatibility with other tree-based ensemble models. Furthermore, while this paper introduces certain metrics and algorithms derived from graph theory, the field offers extensive possibilities for future exploration. In the future, we plan to introduce new tools associated with DPG to enhance the interpretation of tree-based ensemble models.

4.5 Conclusion

In this paper, we introduced DPG as a novel model-specific tool for tree-based ensemble interpretability. DPG is obtained from a trained model and data, ensuring the maintenance of its performance. The concept behind DPG is to convert an opaque-box tree-based ensemble model into an enriched graph. DPG enables graph-based evaluations and the identification of model decisions towards facilitating comparisons between features and their associated values while offering insights into the entire model. In particular, we introduced Betweenness Centrality, Local Reaching Centrality, Community and Constraints as useful metrics and properties towards improving and extending the XAI interpretability approaches. While DPG is still considered an evolving work, its potential is substantial, given the robust underlying theory and the versatility of the tool. Furthermore, the effervescent research on graphs, knowledge graphs, and complex networks might strengthen the possibilities grounded in DPG. As the next step of our current research, we expect to apply DPG to improve global interpretability and to enhance the scalability of the current implementation.

Chapter 5

Enhancing Transparency in the Fruit Supply Chain

In this Chapter, we apply two explainability methods, including the DPG, to a real-world problem and show their effectiveness in improving the interpretability of an AI system for food quality analysis.

A central task in food quality assessment is determining fruit ripeness and detecting possible degradation. This task is often addressed with AI because ripeness relates to physical properties, such as color, and to chemical properties, such as pH. As widely noted, these applications are challenging to interpret, particularly when explanations are limited to a single instance. A global method, such as the DPG, can therefore clarify the overall decision-making process, especially when paired with a local method that explains specific predictions.

Section 5.1 introduces the topic and reviews prior applications of AI and explainability in fruit quality analysis. Section 5.2 describes the datasets and the local method BELLATREX [62], used together with the DPG to explain an RF trained on physicochemical data. Section 5.3 presents the RF results and the explanations produced with both methods, and discusses the insights they provide.

5.1 Introduction

The food supply chain, and particularly the fruit sector, represents an important component of the global economy [202]. Monitoring fruit quality is mandatory to ensure that products available in the market are not only safe and nutritious but also sensorially appealing [180]. Fruit quality has a direct impact on public health, social well-being, and environmental sustainability, influencing responsible production and consumption practices. Meeting consumer expectations is therefore essential to ensure product acceptance, foster brand loyalty, and promote healthy food choices, thereby driving commercial success and supporting long-term sustainability [291].

The fruit industry supply chain is a complex and dynamic network that encompasses production, post-harvest handling, storage, transportation, processing, distribution, and retail. Each stage plays a critical role in ensuring the quality, safety, and sustainability of fruits delivered to consumers, as shown in Figure 5.1, which has been adapted from [180, 269, 291].

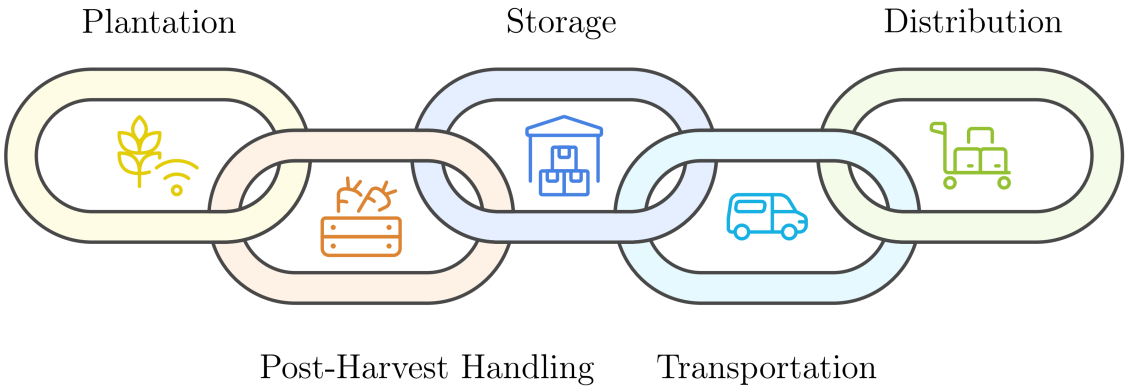


Figure 5.1: Overview of the Fruit Supply Chain as in [180, 269, 291]. Food begins its transformation into a product during cultivation. It is then harvested, processed, stored, and transported to distribution centers for consumer sale.

This supply chain begins with the plantation, where fruits are cultivated on farms or orchards using agricultural practices tailored to local climate and soil conditions. Advanced technologies, such as precision agriculture, optimize resources like water and fertilizer to improve yield and quality [180]. Following harvest, the fruits enter the post-harvest handling stage, where they are sorted, cleaned, and graded based on attributes like size, colour, and ripeness, often assessed using physicochemical and sensory characteristics [306]. The storage phase involves maintaining optimal conditions of temperature, humidity, and gas composition to preserve freshness and extend shelf life. Techniques such as cold storage and modified atmosphere packaging are commonly employed to minimize deterioration [269]. Fruits are then transported to processing facilities or directly to distribution hubs. During transportation, real-time monitoring systems track environmental parameters, ensuring that fruits remain fresh until they reach their destination [200]. At the processing stage, some fruits are transformed into value-added products such as juices, dried fruits, or jams, which involve operations like slicing, peeling, and packaging [291]. The distribution phase moves fruits from wholesalers or producers to retail outlets, including supermarkets, local markets, and online platforms [180].Finally, in the retail and consumer stage, fruits are presented to customers who evaluate them based on appearance, freshness, and nutritional value. Consumer feedback at this stage often informs upstream decisions, closing the loop on the supply chain [291].

Each stage of the fruit supply chain presents opportunities for traceability, optimization, and waste reduction, with emerging technologies like AI and the IoT playing a critical role in enhancing efficiency and transparency. Monitoring the physicochemical parameters of fruits, such as colour, texture, acidity, sugar content, and bioactive compound levels, is essential for maintaining quality, extending shelf life, and meeting consumer expectations. These parameters are critical indicators of ripeness, freshness, and overall quality, and they are influenced by environmental factors like temperature, humidity, and gas composition. Continuous evaluation of these factors during storage and transportation is necessary to detect early signs of physiological changes, such as enzymatic degradation or Moisture Content loss, which can lead to deterioration if unmanaged [237, 306]. Understanding the re-

lationships among these parameters throughout the ripening process and shelf life offers substantial benefits for optimizing storage conditions, ensuring food safety, and preserving both sensory and nutritional qualities. However, transparency in the fruit supply chain remains a major challenge due to the complexity of processes, perishability, and growing consumer demands for quality and sustainability. The lack of standardized systems for monitoring and sharing data on storage conditions, safety, and traceability undermines accountability and efficiency.

AI can tackle challenges in the fruit supply chain by improving monitoring through predictive capabilities [200]. Concrete examples include AI-powered IoT sensors that can track critical parameters, such as temperature and humidity, in real-time, enabling early detection of issues like Moisture Content loss or enzymatic degradation. Furthermore, ML models can analyze data to predict quality changes (e.g., [75, 232]) and recommend optimal storage conditions (e.g., [67]). At the same time, computer vision systems and DL methods assess visual quality indicators to detect spoilage (e.g., [77, 122, 264]). Due to their versatility, ease of interpretation, and the wide range of XAI techniques available to analyze their decision-making processes, tree-based ensemble learning models are highly effective. When integrated with XAI, these models enhance both robustness and interpretability by leveraging the strengths of multiple models and providing valuable, actionable insights. This combined approach not only helps identify biases or errors in the data and models but also supports more accurate, reliable, and informed decision-making, which is especially important in critical areas such as public health and food safety [76]. However, in many studies, the potential to combine different explanations for a more comprehensive and effective solution—such as integrating local and global approaches into a "glocal" method—is not fully explored.

In this project, we explore integrating tree-based ensemble learning and XAI into the fruit industry supply chain. By leveraging the physicochemical properties of fruits, we use the RF algorithm to predict quality and identify key factors affecting sensory characteristics. We apply XAI methods—DPG and BELLATREX—to provide both global and local explanations for three real-world case studies: carambola, pitaya, and papaya. These approaches enhance interpretability, offering valuable insights into the decision-making process across the supply chain.

DPG, a global technique, provides a visual representation of the decision-making process by modeling decision predicates and their relationships, helping users understand how inputs lead to outcomes in complex models. BELLATREX, a local technique, generates accurate, human-readable rules for individual predictions, enhancing the interpretability of tree-based ensemble models.

Our primary contribution is demonstrating the benefits of combining tree-based ensemble methods with XAI techniques, highlighting how they can be integrated into a glocal approach. This combination improves supply chain interpretability, offering distinct advantages. The findings are particularly relevant for post-harvest quality assessment, where evaluating ripeness, nutritional quality, and sensory appeal is essential for optimizing storage, transportation, and market readiness.

Table 5.1: Physicochemical features used in the Carambola, Pitaya, and Papaya Datasets. These attributes were measured to characterize fruit condition and support the prediction of post-harvest quality levels.

Feature	Definition
Ascorbic Acidity	Refers to the vitamin C content in fruit.
pH	Indicates the acidity or alkalinity of the fruit.
Titratable Acidity	Represents the total acidity, primarily due to organic acids like citric, malic, and Ascorbic Acidity.
Soluble Solids (Soluble Solids Concentration)	Typically measured in °Brix, this indicates sugar content and is linked to sweetness.
Moisture Content	Reflects the water content of the fruit.
Phenolic Content	Relates to antioxidant compounds in the fruit.
Pulp Firmness	A measure of texture and ripeness.
Carotenoids	Pigments that contribute to yellow-orange coloration and have antioxidant properties.
L*	A colorimetric parameter representing lightness (ranging from 0 (black) to 100 (white)).
a*	Measures red-green color axis with positive values representing red and negative values representing green.
b*	Represents the blue-yellow axis with positive values representing yellow and negative values representing blue.
C (Chroma)	Calculated from a* and b*, it reflects color saturation or intensity.
Hue (Hue Angle)	The angle in the a-b color space that corresponds to the perceived color tone.

5.2 Materials and Methods

5.2.1 Fruit Quality Studies

To demonstrate the benefits of using AI to improve the fruit supply chain, we employed three publicly available datasets previously studied in the literature. These datasets consist of numerical features and are suitable for predicting the post-harvest quality level of various fruits. Specifically, the work by [59] focuses on a dataset describing the characteristics of carambola; [206] utilizes a dataset related to papaya; and [55] presents a study on the characteristics of pitaya. As reported by the respective authors, the fruits were acquired from local markets and producers in Campinas, São Paulo, Brazil, and sampled across different batches and seasons to ensure a wide range of variability. Moreover, multiple measurements were collected for each sample, resulting in datasets containing a variety of both physical and chemical features. These were designed to characterize the condition of the fruits and to associate them with a quality level that typically degrades over time. The physicochemical features evaluated are listed in the Table 5.1.

In this subsection, we provide a detailed description of the datasets and the various features collected from the fruit samples. Table 5.2 summarizes the specific attributes measured in each dataset, which were used to characterize the samples.

Table 5.2: Presence of each physicochemical feature in the Carambola, Pitaya, and Papaya Datasets. The table indicates which attributes were measured and included in the respective datasets used for fruit quality analysis.

Attribute	**Carambola** [59]	**Pitaya** [55]	**Papaya** [206]
Ascorbic Acidity	X		X
pH	X	X	
Titratable Acidity	X	X	
Soluble Solids	X	X	X
Moisture Content	X	X	
Phenolic Content		X	
Pulp Firmness			X
Carotenoids			X
L*	X		X
a*	X		X
b*	X		X
C	X		X
Hue	X		X

Carambola Dataset

A total of 177 fruit samples from the B17 and B10 varieties—the most commercially cultivated types—were collected from the Centrais de Abastecimento (CEASA) Supply Center on multiple dates. B17 fruits are typically sweeter, less acidic, and possess superior edible quality compared to B10, although they have a shorter shelf life [1]. All samples were disinfected using a chlorine-based solution ($0.2\,\mathrm{g\,L^{-1}}$) and stored at $23\,^{\circ}\mathrm{C}$ under controlled conditions prior to analysis. Additional physicochemical properties—ascorbic acididity, pH, titratable acidity, soluble solids concentration, and moisture content—were also measured and incorporated into the dataset. To evaluate ripening, samples were classified into four Maturity Stage (MS) based on the hue angle measured with a colorimeter. The first stage (MS1) corresponded to green fruits with hue $> 100°$; MS2 to green/yellow fruits with $92° < \text{hue} \leq 100°$; MS3 to yellow fruits with $83° < \text{hue} \leq 92°$; and MS4 to yellow/orange fruits with hue $\leq 83°$. The dataset exhibits class balance, containing approximately 45 samples in each category.

Papaya Dataset

The dataset consisted of 57 papaya samples sourced from a retail market. For each fruit, two color images, one for each side, were acquired to document external characteristics. In addition, several physicochemical attributes were measured, including pulp firmness, pH, soluble solids content, total carotenoids, and ascorbic acid concentration. Based on pulp firmness, samples were classified into three maturity stages, following the criteria established by [206]. The first stage (MS1) included fruits with pulp firmness greater than 33 N; the second stage (MS2) included those with pulp firmness between 20 N and 33 N; and the last stage (MS3) comprised fruits

with pulp firmness below 20 N. Pulp firmness is considered a critical indicator of ripeness, as papayas with pulp firmness below 20 N are typically regarded as fully edible [27, 129, 206]. Differences in the number of samples across maturity stages were a result of the initial visual classification used during sample selection. Class distribution in the dataset is uniform, with each class represented by approximately 20 samples.

Pitaya Dataset

A total of 140 red-fleshed pitaya fruits were collected immediately after harvest from a local producer and divided into several batches. Thirty samples were analyzed on the day of harvest to establish baseline quality parameters. The remaining 110 fruits were stored under two temperature conditions (15 °C and 25 °C) and analyzed after 7, 14, 21, and 25 days to monitor changes in quality over time. The analyses included measurements of total phenolic content, soluble solids concentration, pH, titratable acidity, and moisture content. It is important to note that, unlike the other two datasets, fruit maturity in this case is directly linked to the passage of time. Consequently, the traditional maturity stages are replaced by a temporal indicator: the number of days elapsed since harvest, which serves as the defining criterion for maturity level in this context. Each class in the dataset is represented by approximately 30 samples, resulting in a balanced distribution.

5.2.2 Machine Learning Methods

An RF classifier was trained on each of the three datasets to perform a classification task aimed at predicting fruit maturity levels. Each RF model was composed of 100 tree-based learners and was trained on 70% of the data, with the remaining 30% reserved for testing.

Model performance was evaluated using three key metrics. Accuracy measured the proportion of correctly classified instances relative to the total number of instances in the test set. F1-score was used to provide a balanced assessment by incorporating both precision and recall. Furthermore, confusion matrices were examined to offer a detailed understanding of classification performance by identifying the nature and frequency of the model's misclassifications.

5.2.3 Explainable Artificial Intelligence Methods

Our work takes advantage of two recent XAI methods, DPG [10], introduced in Chapter 4, and BELLATREX [62], to provide a comprehensive understanding of the decision-making processes of the RF models. The combination of these methods not only reveals FI but also uncovers additional insights that contribute to a more transparent interpretation of the models' behavior.

BELLATREX

BELLATREX is a local, model-specific method designed to explain predictions made by a RF model. It condenses the decisions of individual tree learners into a coherent

Table 5.3: Confusion matrix showing the classification performance of the RF model on the Carambola Dataset using four stages (MS1 to MS4).

	Prediction			
Ground truth	MS1	MS2	MS3	MS4
MS1	16	0	0	0
MS2	1	9	0	0
MS3	0	0	14	0
MS4	0	0	0	14

explanation for how the model classifies a specific instance. The process consists of four key steps. First, for each input, a set of rules is extracted from the RF, capturing the decision paths of the trees that contribute to the model's output. These rules are then transformed into high-dimensional vector representations that encode both the structural conditions of each rule and their relevance to the specific classification. Next, dimensionality reduction is applied which projects the vectors into a lower-dimensional space while preserving their relative similarities. In this simplified space, similar rules are grouped into clusters, each reflecting a consistent reasoning pattern used by the model. From each cluster, BELLATREX selects a representative set of rules—usually the one nearest to the cluster's centroid—which serves as a prototype summarizing that group's rationale. These prototype rules form a concise and interpretable explanation of the RF's decision, highlighting the key features and logic that influenced the classification outcome for the specific instance.

5.3 Results

5.3.1 Carambola Dataset

The RF model with 100 trees, trained on the Carambola dataset, achieved excellent results, with an accuracy of 98.15% and an F1-score of 98.13%, correctly classifying 53 out of 54 test samples, as shown in Table 5.3. The result confirmed the effectiveness of RF in assigning ripeness level to carambola samples based on various physicochemical characteristics.

Table 5.4 illustrates the explanation provided by BELLATREX when applied to two randomly selected samples from the Carambola test set. In the first example, BELLATREX classifies the sample as belonging to the MS2 class, aligning with the prediction made by the RF model. This classification is based on three features: the characteristics L*, a*, and hue. Specifically, BELLATREX projects the vector representations of selected rules into two dimensions, grouping them into a single cluster. The rule closest to the cluster's centre is then chosen as the final representative. The prediction for the second example follows a similar process, with both BELLATREX and RF classifying it as part of the MS3 class. However, in this case, BELLATREX incorporates additional chemical features, namely titratable acidity and ascorbic acidity, alongside the physical characteristics.

Table 5.5 presents the most important predicates of the DPG ranked by their

Table 5.4: Explanation provided by BELLATREX technique applied over two randomly selected samples of the Carambola test set. The first belongs to class MS2 and the second belongs to class MS3.

Extracted rule sample 67	weight=1.00	Initial estimate = (0.27, 0.24, 0.24, 0.26)
Instance	Split test	Prediction
L* = 43.80	L* > 39.22	(0.29, 0.34, 0.34, 0.03)
Hue = 98.97	Hue > 91.77	(0.46, 0.55, 0.00, 0.00)
a* = -5.27	a* > -6.11	(0.00, **1.00**, 0.00, 0.00)
Extracted rule sample 119	weight=1.00	Initial estimate = (0.29, 0.16, 0.31, 0.24)
Instance	Split test	Prediction
Titratable Acidity = 0.30	Titratable Acidity ≤ 0.44	(0.07, 0.19, 0.43, 0.31)
b* = 39.68	b* > 32.42	(0.03, 0.07, 0.51, 0.39)
Ascorbic Acidity = 0.51	Ascorbic Acidity > 0.28	(0.00, 0.09, 0.30, 0.61)
Hue = 86.31	Hue > 83.15	(0.00, 0.24, 0.77, 0.00)
L* = 47.70	L* ≤ 48.40	(0.00, 0.08, 0.92, 0.00)
a* = 2.56	a* > -0.92	(0.00, 0.00, **1.00**, 0.00)

Table 5.5: BC for DPG predicates of the RF trained on Carambola Dataset.

Predicate	**BC**
a* ≤ -1.13	0.0487
Hue ≤ 91.94	0.0421
Hue > 83.16	0.0381
Hue > 91.94	0.0365
a* > -1.13	0.0334
Titratable Acidity ≤ 0.44	0.0303
b* > 34.19	0.0276
Hue ≤ 99.94	0.0246
Hue > 83.0	0.0238
a* ≤ 4.26	0.0236

Table 5.6: LRC for DPG predicates of the RF trained on Carambola Dataset.

Predicate	**LRC**
a* > -1.03	1.1615
pH > 3.07	1.1497
Hue ≤ 91.94	1.1248
Hue > 83.0	1.1042
a* > -5.93	1.0936
a* > -5.84	1.0916
Hue ≤ 100.53	1.0890
Titratable Acidity ≤ 0.51	1.0728
a* ≤ 3.76	1.0686
Soluble Solids > 9.39	1.0676

BC values. Several predicates emerge as particularly noteworthy. Notably, the predicates with the highest indices correspond to the physical features a* and hue, associated with values of -1.13 and 91.94, respectively. These predicates reflect critical decisions made by a significant number of tree-based learners. Table 5.6 highlights the ten most significant predicates based on LRC: in addition to the previously mentioned physical features (a* and hue), the predicate involving the chemical feature pH is identified as particularly influential.

The comparison between the local explanations provided by BELLATREX and the global insights derived from the DPG underscores the relevant role of physical characteristics in classifying carambola fruits. Notably, parameters related to colour are particularly informative, as they reflect the natural progression of ripening. The colour change observed during ripening is closely linked to underlying physicochemical transformations. As the fruit matures, chlorophyll, the pigment

Table 5.7: Confusion matrix depicting the RF model's performance in classifying three maturity stages (MS1 to MS3) of the Papaya Dataset.

	Prediction		
Ground truth	MS1	MS2	MS3
MS1	16	0	0
MS2	1	9	0
MS3	0	0	14

responsible for its green hue, degrades, revealing other pigments such as carotenoids and anthocyanins. The synthesis of carotenoids during this stage contributes to the development of the yellow and orange tones characteristic of ripe carambola. The identification of the hue as the most influential feature is consistent with its role in defining the ripeness levels of carambola fruits within the dataset. This alignment confirms that the RF model effectively based its classification on the most relevant attribute, thereby reinforcing the validity of its decision-making process.

5.3.2 Papaya Dataset

With a configuration of 100 trees, the proposed RF model demonstrated strong performance on the Papaya dataset. It achieved an accuracy of 97.50% and an F1-score of 97.53%. As shown in the confusion matrix (Table 5.7), the model correctly classified 39 out of 40 test samples, confirming its ability to accurately assign ripeness levels based on physicochemical attributes.

Table 5.8: Explanation provided by BELLATREX technique applied over two randomly selected samples of the Papaya test set. The first belongs to class MS3 and the second belongs to class MS1.

Extracted rule sample 54	weight=1.00	Initial estimate = (0.457, 0.261, 0.283)
Instance	Split test	Prediction
Ascorbic Acidity = 131.09	Ascorbic Acidity $>$ 121.23	(0.375, 0.000, 0.625)
Hue = 88.60	Hue $\leq$ 93.74	(0.000, 0.000, **1.000**)
Extracted rule sample 5	weight=1.00	Initial estimate = (0.500, 0.283, 0.217)
Instance	Split test	Prediction
pH = 5.90	pH $>$ 5.69	(0.605, 0.263, 0.132)
b* = 38.41	b* $\leq$ 47.24	(0.821, 0.179, 0.000)
L* = 57.06	L* $\leq$ 60.23	(0.920, 0.080, 0.000)
pH = 5.90	pH $>$ 5.81	(**1.000**, 0.000, 0.000)

The application of BELLATREX to two randomly selected examples, the first one belonging to the MS3 class and the second to the MS1 class highlights two different behaviours (Table 5.8). Even in these instances, the vector representations of the selected rules were projected onto two dimensions and grouped into a single cluster. However, we can observe some differences in the exploited characteristics

Table 5.9: BC for DPG predicates of the RF trained on Papaya Dataset.

Predicate	**BC**
Ascorbic Acidity $\leq$ 119.12	0.0787
b* $\leq$ 46.01	0.0705
Hue $\leq$ 92.21	0.0636
L* $\leq$ 60.1	0.0556
Hue $\leq$ 90.21	0.0501
Ascorbic Acidity $\leq$ inf	0.0419
Hue $\leq$ 93.24	0.0412
Hue $>$ 93.06	0.0364
Pulp Firmness $>$ 3.14	0.0265
C $\leq$ 49.45	0.0254

Table 5.10: LRC for DPG predicates of the RF trained on Papaya Dataset.

Predicate	**LRC**
Hue $>$ 92.21	0.7892
a* $\leq$ -0.66	0.7732
L* $\leq$ 59.94	0.7664
a* $\leq$ -0.86	0.7614
L* $\leq$ 60.23	0.7452
pH $>$ 5.69	0.7426
C $\leq$ 44.14	0.7419
a* $\leq$ -0.72	0.7401
Hue $>$ 92.39	0.7298
L* $\leq$ 58.2	0.7251

for the two examples. In particular, the first instance is classified based on ascorbic acidity and hue, utilizing both physical and chemical characteristics. In contrast, the second instance relies on different features, specifically pH, and the b* and L* values. This highlights that entirely distinct features are employed for different instances. Moreover, it is noteworthy that the use of ascorbic acidity as a feature appears peculiar, as there is no evident correlation between this feature and the classes. Additionally, certain samples exhibit missing values for this feature.

In Table 5.9 we can observe the most important predicates concerning their BC values. Predicates involving ascorbic acidity are extensively utilized by the tree-based learners in the model. Contrary to the explanation provided by BELLATREX, the prominence of this feature can be attributed to the presence of missing values, leading to a feature-value association, specifically ascorbic acidity with 119.12, which becomes predominant. The remaining predicates presented in the table primarily involve physical characteristics associated with colour. In the Table 5.10, which ranks the best predicates based on LRC values, it is evident that apart from pH and C, the predicates predominantly include physical characteristics.

As with the Carambola dataset, the application of XAI techniques reveals that papaya colour undergoes significant changes during ripening, primarily due to chlorophyll degradation and increased carotenoid accumulation. XAI analysis also highlights the role of pH and ascorbic acidity in influencing colour stability and intensity. During ripening, the pH of papaya may shift due to a reduction in organic acids and an increase in sugar content. These chemical changes not only enhance flavour but also affect colour perception, contributing to the characteristic yellow appearance of ripe papaya. However, these findings highlight an important consideration. In the dataset, ripeness levels were defined based on pulp firmness; nevertheless, this feature did not rank among the most influential in the RF model's classification process. While the model's decision-making is grounded in physiologically relevant attributes such as colour, it does not fully align with the expert-defined criteria, which identified pulp firmness as the primary indicator of ripeness.

Table 5.11: Confusion matrix of the RF model applied to the Pitaya Dataset, evaluated across five postharvest time points (0, 7, 14, 21, and 25 days).

	Prediction				
Ground truth	0 days	7 days	14 days	21 days	25 days
0 days	9	2	0	0	0
7 days	1	6	0	0	0
14 days	0	0	9	1	0
21 days	0	0	4	5	0
25 days	0	0	0	0	5

Table 5.12: Explanation provided by BELLATREX technique applied over two randomly selected samples of the Pitaya test set. The first belongs to the class of samples analyzed after 21 days and the second belongs to the samples analyzed after 14 days.

Extracted rule sample 119	weight=1.00	Initial estimate = (0.22, 0.21, 0.25, 0.17, 0.14)
Instance	Split test	Prediction
pH = 5.10	pH > 4.85	(0.00, 0.00, 0.44, 0.31, 0.26)
Titratable Acidity = 0.14	Titratable Acidity > 0.12	(0.00, 0.00, 0.56, 0.42, 0.02)
Soluble Solids = 10.80	Soluble Solids ≤ 12.70	(0.00, 0.00, 0.47, 0.50, 0.03)
Titratable Acidity = 0.14	Titratable Acidity ≤ 0.15	(0.00, 0.00, 0.20, 0.73, 0.07)
Soluble Solids = 10.80	Soluble Solids > 9.10	(0.00, 0.00, 0.00, 0.92, 0.08)
Moisture Content = 87.50	Moisture Content > 84.66	(0.00, 0.00, 0.00, **1.00**, 0.00)
Extracted rule sample 31	weight=1.00	Initial estimate = (0.21, 0.19, 0.33, 0.14, 0.12)
Instance	Split test	Prediction
pH = 4.50	pH > 4.25	(0.00, 0.24, 0.42, 0.18, 0.16)
Moisture Content = 89.70	Moisture Content ≤ 89.71	(0.00, 0.18, 0.47, 0.21, 0.15)
Phenolic Content = 0.69	Phenolic Content > 0.62	(0.00, 0.20, 0.53, 0.23, 0.05)
Titratable Acidity = 0.22	Titratable Acidity > 0.19	(0.00, 0.83, 0.17, 0.00, 0.00)
pH = 4.50	pH ≤ 4.75	(0.00, **1.00**, 0.00, 0.00, 0.00)

5.3.3 Pitaya Dataset

On the Pitaya dataset, the RF model with 100 trees achieved an accuracy of 80.95% and an F1-score of 80.61%. As reported in the confusion matrix (Table 5.11), the model correctly classified 34 out of 42 test samples, demonstrating reasonable effectiveness in predicting ripeness levels based on the available chemical features.

In Table 5.12, the application of BELLATREX is demonstrated on two randomly selected examples: the first belonging to the 21-day class and the second to the 14-day class. Unlike the studies on the Carambola and Papaya Datasets, the Pitaya Dataset consists solely of chemical features. As in previous cases, the vector representations of the selected rules were projected onto two dimensions and grouped into a single cluster. The two rules presented in the table exhibit similar characteristics for each example. For instance, pH emerges as a particularly significant classification feature in the first case, as it effectively eliminates certain classes when

Table 5.13: BC for DPG predicates of the RF trained on Pitaya Dataset.

Predicate	**BC**
Phenolic Content $>$ 0.72	0.1646
Titratable Acidity $\leq$ 0.18	0.1042
Titratable Acidity $>$ 0.11	0.0993
Phenolic Content $>$ 0.62	0.0868
pH $>$ 4.85	0.0854
Phenolic Content $>$ 0.6	0.0849
pH $\leq$ 4.85	0.0686
pH $\leq$ 5.35	0.0660
pH $>$ 5.16	0.0605
Phenolic Content $\leq$ 0.72	0.0565

Table 5.14: LRC for DPG predicates of the RF trained on Pitaya Dataset.

Predicate	**LRC**
pH $\leq$ 4.85	2.1466
Soluble Solids $>$ 9.47	2.0335
pH $>$ 4.75	1.9887
pH $>$ 4.85	1.9483
Moisture Content $>$ 84.77	1.9477
Soluble Solids $\leq$ 13.18	1.9368
Soluble Solids $>$ 10.42	1.9336
Moisture Content $>$ 85.33	1.8730
Moisture Content $>$ 82.4	1.8596
Phenolic Content $>$ 0.62	1.8556

involved in models' decisions.

Table 5.13 presents the most important predicates of the DPG, ranked by their BC values. The predicates with the highest indices correspond to the features titratable acidity, phenolic content, and pH. Conversely, Table 5.14, which highlights the most significant predicates based on LRC, identifies pH, soluble solids, and moisture content as the most prominent features. A comparison of the two XAI methods reveals that certain features play a predominant role in classifying the ripeness levels of pitaya. While many features are identified as more or less relevant, pH clearly emerges as the most influential variable in the model's final classification. In particular, the pH value of 4.85 is consistently highlighted by both DPG and BELLATREX, marking it as a critical feature–value pair for classifying a large portion of the dataset samples. The pH of pitaya changes during its shelf life due to biochemical, microbiological, and physiological processes, making it a reliable indicator of quality. Enzymes within the fruit contribute to the degradation of organic acids over time, leading to a reduction in acidity, especially as the fruit ripens. The central role of pH, both as the most frequently used and most influential feature in the model, confirms the effective functioning of the RF model in accurately classifying pitaya fruit based on meaningful indicators.

5.4 Conclusion

This study applied two XAI techniques, namely DPG and the BELLATREX, to improve the interpretability of AI models for fruit ripeness classification. Using real-world datasets (Carambola, Pitaya, and Papaya), we demonstrated the effectiveness of these methods in providing both global and local (glocal) explanations.

The proposed RF models successfully classified the maturity stages of all three fruits based on physicochemical features, achieving near-perfect accuracy for Carambola and Papaya, and solid performance for Pitaya. The use of XAI techniques confirmed model reliability by offering clear and interpretable insights into FI and decision logic.

By embedding explainability into AI-driven decisions, this approach supports critical supply chain functions, including quality control, traceability, and compli-

ance with food safety standards. For example, in the Papaya dataset, the model's reliance on visual features such as hue and physical attributes like pulp firmness may direct stakeholders to actionable decisions related to packaging and logistics.

Moreover, understanding ripeness stages enables more efficient fruit utilization. For instance, ripe fruits can be prioritized for immediate distribution, while less mature ones are suitable for processing. Interpretable models enable stakeholders to make informed and proactive choices, supporting transparency and accountability.

In sum, this work highlights the value of combining AI with "glocal" XAI methods to promote interpretability, reliability, and ethical decision-making in high-stakes domains such as food safety and public health.

Future work could extend this framework to integrate IoT-enabled real-time monitoring systems, enabling more proactive and dynamic quality control. Additionally, the methodology could be generalized to other perishable goods, broadening its impact on global food supply chains.

Chapter 6

Extending Decision Predicate Graphs for Isolation Forest

In this Chapter, we extend DPG to introduce a novel, model-specific XAI technique that provides a global explanation of iForest.

An ML pipeline encompasses the entire process from raw data to model deployment and final outputs. These pipelines consist of several interconnected stages. The initial stages focus on understanding the dataset and enhancing data quality, as the integrity of the inputs significantly impacts the model's performance in later stages. This preliminary work is referred to as preprocessing and includes various techniques, such as cleaning duplicate records, correcting inconsistent labels or units, and handling missing data and outliers. Thorough preprocessing minimizes noise and bias, leading to more stable model performance. This phase often utilizes AI methods that may not be entirely transparent. For example, the iForest algorithm is commonly used to identify outliers in a dataset. As a tree-based ensemble method, its decision-making process can be clarified through DPG, which helps explain how outliers are detected.

Section 6.1 introduces the need for transparency in predictive models and preprocessing methods such as outlier detection with iForest. In Section 6.2, we review existing methods that provide explanations for the iForest algorithm. Section 6.3 formalizes our new technique, detailing its construction and the insights that can be derived from the resulting explanation. Section 6.4 presents the method through three case studies, illustrating its functionality and benefits. In Section 6.5 we discuss the limitations of the approach and outline possible improvements.

The content of this Chapter has been published in [39]

6.1 Introduction

Most current XAI techniques primarily focus on elucidating predictive models, often overlooking the need to address the entire data processing pipeline. This partial focus can result in incomplete explanations of the context, potentially obscuring critical aspects of data handling and preprocessing. As Lipton [142] argues, a holistic approach to explainability is essential for the credibility and utility of ML solutions. Similarly, authors advocate for a shift towards transparent ML ecosystems, where ev-

ery pipeline component, from data preprocessing to model decision-making, is made transparent [107, 275]. More robust, trustworthy explanations can be constructed by ensuring XAI techniques encompass the entire pipeline. Data preparation and transformation models before training a predictive model demand clarity equal to the last one for several reasons, including transparency, reliability, and regulatory requirements [253]. Firstly, transparency in preprocessing enhances the understandability of the data manipulations that occur before model training [255]. By understanding how data is cleaned, normalized, and selected during preprocessing, users can identify potential sources of bias or errors that might affect the model's performance. Furthermore, this process enables the detection and mitigation of data-acquisition issues, such as systematic errors and noise, and supports enhancements to the overall system pipeline. Finally, clear documentation and explanation of all stages of data handling, including preprocessing, ensure compliance with these regulations and promote trust and reliability [297].

Among many preprocessing algorithms, iForest [143] stands out for its straightforward approach and its effectiveness in swiftly handling outliers in high-dimensional data. However, the core mechanism of iForest, which relies on random selection of features and split points to isolate outliers, introduces stochasticity that can sometimes yield ambiguous or non-intuitive results [125]. Consequently, providing explanations for iForest's decisions is essential, as it allows users to understand and trust the logic behind outlier identification, particularly when dealing with complex datasets. These explanations not only help validate the outliers detected by iForest but also aid in fine-tuning the model by revealing potential biases or errors introduced by the randomness in the selection process [187]. SHAP [148] is currently used to explain the behaviour of the iForest model by providing insights into how features influence its predictions. In contrast, the Depth-based Isolation Forest Feature Importance (DIFFI) [35] method employs a tailored approach that leverages iForest's internal structure to compute FI. However, both methods provide local explanations that use a FI vector to illustrate the model's decision-making process for identifying individual samples. While effective, these approaches primarily focus on feature-level contributions, neglecting the structural and logical complexities of the iForest ensemble.

To overcome the limitation of providing only a vector of FI, we propose a novel DPG-based method to elucidate the logic and intrinsic properties of the iForest ensemble. Building on the principles of the DPG technique, our method converts the iForest model into a graph, enabling us to exploit its structural properties and leverage established mathematical theories to elucidate the outlier-detection process. The proposed method is global, as it explains the entire decision-making process of the iForest model, revealing general patterns and the feature interactions that drive the whole model's logic. This approach provides a mixed-type explanation, as done in other research [155, 224], by integrating a visual representation of the model's decision-making process with a new quantitative metric, the IOP-Score, which assesses each feature's contribution to outlier detection. By extracting relationships and decision paths within the ensemble, our method enhances model interpretability and delivers actionable insights into its internal mechanisms, surpassing traditional

explanation techniques. This work contributes in the following ways:

- Comprehensive global explanation of iForest: we propose a method to explain the iForest model, including details on feature boundaries for both inlier and outlier samples.
- The IOP-Score: a novel metric that quantifies a node's tendency to propagate toward either the inliers or outliers to enhance interpretability by distinguishing discriminative from neutral predicates in the iForest.
- Graph-based interpretability: by integrating DPG, we introduce a graph-based structure that models the isolation logic, such as feature influence on isolation depth and decision paths, enabling a detailed understanding of the detection process.

The results are derived from synthetic and well-established datasets to demonstrate the method's potential. However, we emphasize that the approach is generalizable, indicating its broad applicability across various related scenarios.

6.2 Literature Review

The literature presents several post-hoc XAI methods designed to interpret the iForest model. Post-hoc XAI methods are applied after training to provide interpretability without altering the model's internal structure, thereby preserving its performance. According to Speith [254], we can distinguish the model-agnostic XAI methods, such as SHAP [148], which can be applied independently of the underlying model, from the model-specific method, tailored for specific models or model classes.

Considering proposals using SHAP, Rachwał et al. [215] proposed an improved iForest algorithm that dynamically excludes attributes based on SHAP indices, resulting in enhanced prediction accuracy and better feature selection. In their approach, SHAP values are used to quantify the importance of each feature, and models are iteratively trained with one feature excluded at a time. The final anomaly score of iForest is computed as a weighted average of these models' anomaly scores, where the weights are derived from the absolute SHAP values, prioritizing features with higher SHAP values and reducing the influence of less relevant ones.

Liu and Aldrich [144] introduced the iForest-RF-SHAP framework, a novel approach for anomaly detection and explanation in coal data, which combines iForest, RF, and SHAP. This framework outperformed traditional methods, such as principal component analysis, while offering detailed insights into variable contributions.

Finally, there are some XAI methods that are model-specific for iForest and are introduced in [7, 35, 125].

Kartha et al. [125] developed a method specifically designed to explain iForest outliers predictions by assigning a vector of FI weights to each attribute, indicating its contribution to the anomaly detection process. These weights are computed by analyzing how much each attribute contributes to isolating a data point within the iForest trees, with higher weights associated with shorter path lengths. The result is an explanation vector that reflects the relative importance of each feature in determining the anomaly score.

Arcudi et al. [7] introduced Extended Isolation Forest Feature Importance (ExIFFI), a method designed to deliver global and local explanations for iForest. ExIFFI uses FI metrics to explain outliers detection comprehensively, offering a detailed perspective on how individual features contribute to the model's predictions. The FI metrics are computed by analyzing the projections of the hyperplane's normal vector at each node in the isolation trees and weighting them based on the degree of imbalance in the data split, favoring nodes where the sample falls into the smaller partition, thus attributing greater importance to features that isolate outliers more effectively.

Carletti et al. [35] presented DIFFI, a method tailored for iForest. DIFFI provides global and local interpretability by analyzing how features influence the depth at which outliers are isolated in the decision trees. This method explains the anomaly detection process and enables unsupervised feature selection, a valuable tool for handling high-dimensional data in outliers detection problems.

Despite the advancements in explaining iForest models using methods like SHAP, ExIFFI, and DIFFI, a significant gap remains in providing detailed interpretability regarding the values, intervals, and specific characteristics of inliers alongside outliers. This lack of explanation motivates the development of the DPG-based method, which aims to address these limitations.

6.3 DPG-based explanation for Isolation Forest

We propose a novel post-hoc method based on DPG, a model-specific XAI technique designed to understand the decision-making process of the iForest model. An overview of our proposed approach can be seen in Figure 6.1. Subsection 6.3.1 provides an in-depth explanation of each step.

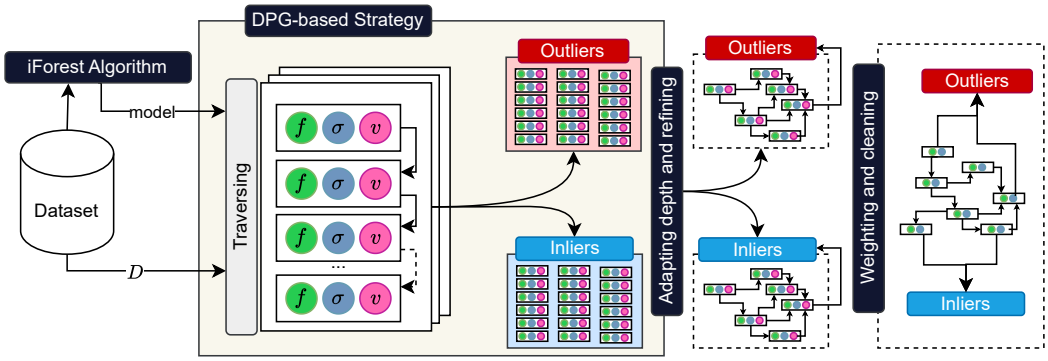


Figure 6.1: Overview of the proposed approach: iForest DPG representation. Predicates are represented as triples (f, σ, v) and are color-coded (green, blue, and pink). The output is the DPG showing the decision-making process of the iForest model, highlighting the two classes into which the data is classified: "Inliers" or "Outliers".

DPG converts the tree-based ensemble model into a graph and uses that structure to provide a global explanation grounded in graph metrics while also visualizing the decision process. We extend this idea to iForest by transforming the model to

examine how the dataset is processed and how outliers are identified. In contrast to FI methods that produce vector rankings, the proposed approach exploits the graph to reveal decision paths, feature interactions, and hierarchical dependencies. This approach improves interpretability by capturing the inner workings of the iForest model, highlighting the key decisions and the most frequently used features for outlier identification. It provides a comprehensive global mixed-type explanation by combining a visual representation of the model's entire decision-making process with a metric that quantifies the relevance of each feature.

In this section, we detail the construction of our technique and present an in-depth explanation of its components. Additionally, we discuss the necessity of this technique, its advantages, and the key insights it offers into the model's behavior.

6.3.1 Proposed Global Explainability

Applying iForest: To construct the explanation, we begin with the iForest model trained on the dataset. The objective is to comprehend the model's decision-making process and identify features differentiating inliers from outliers. The model's output consists of the observations classified as outliers. These observations are assigned labels: "Outlier" if the model classifies them as such, and "Inlier" otherwise.

DPG-based strategy: Following the DPG proposal, we examine the internal nodes of each tree-based learner in iForest, which contain the dataset's split rules used to construct the predicates defined in DPG. These predicates are represented as triples (f, σ, v), where the sign (σ) can be either $>$ or $\leq$. Subsequently, each training sample traverses each tree. We identify all predicate lists satisfied by the samples in each tree-based learner. Each list is then extended by appending the label previously assigned to the observations: "Outlier" if the list results from an outlier's traversal of the tree, and "Inlier" otherwise. As a result, each observation is associated with a set of predicate lists.

Adapting to an iForest DPG: To align with the principles of iForest, which classifies observations that reach the maximum tree depth as inliers, we eliminate all predicate lists that exceed the trees' maximum depth from the outlier sets. This step is crucial because iForest identifies outliers based on their early isolation, i.e., when an observation becomes separated in a leaf before reaching the maximum depth. Since observations that reach this depth may not be truly isolated or may not exhibit outlier characteristics, their removal prevents ambiguity that could lead to their misclassification as inliers.

After generating the predicate lists, we further refine them by removing the values (v) from each predicate triple, resulting in pairs of the form (f, σ). From now on, we will refer to these pairs as predicates.

This abstraction is necessary because iForest selects the split value (v) randomly at each node and for each tree. As a result, the exact triples (f, σ, v) are typically unique to individual trees and are not shared or reused across trees. Aggregating predicates at the level of (f, σ, v) would therefore hinder cross-tree analysis and reduce the generalizability of the method. By focusing on the feature and direction of

the split only, we retain a meaningful and aggregable representation of the isolation patterns across trees.

Weighting iForest DPG: Using the predicate lists, we construct a weighted directed graph that represents the entire model. The predicates serve as the nodes of the graph. A node is connected to another if, within the predicate lists, the predicate in the first node is immediately followed by the predicate in the second node. This ensures that the connection represents the sequential order in which the predicates are satisfied during a decision tree's traversal. The graph's edges represent the frequency with which the pair of predicates stored in the connected nodes appears consecutively in the predicate lists, with the order preserved. The resulting graph shows two classes: "Outlier" and "Inlier", with their respective predicates distinguished by their frequency and position within the model logic.

Cleaning iForest DPG: We can observe that when there is a significant imbalance between the number of outliers and inliers, adjusting the frequency calculation becomes necessary to ensure a fair comparison between the two classes. Predicates satisfied by outliers appear considerably less frequently than those satisfied by inliers. Consequently, identifying the distinctive predicates of each class becomes particularly challenging due to the low frequency of those associated with outliers. We, therefore, introduce a weighting system for the frequencies. For each dataset instance that traverses the model, the transition between two consecutive predicates contributes differently depending on the class assigned to the data point. If the instance is classified as an outlier, its contribution to the frequency is multiplied by a weight w_o. Otherwise, its contribution is multiplied by a weight w_i. The weights are defined as:

$$w_o = \frac{N_o + N_i}{N_o}, \quad w_i = \frac{N_o + N_i}{N_i}, \tag{6.1}$$

where N_o and N_i denote the number of outliers and inliers in the dataset, respectively. We can, therefore, state that the transition between two consecutive predicates satisfied by an outlier has a weighted frequency equal to w_o, while that satisfied by an inlier has a weighted frequency equal to w_i. The weight of an edge is calculated as the sum of these weighted frequencies; for brevity, we refer to this sum as the *weighted frequency of the edge*.

Towards Explanation. Once the graph is constructed, we define a new metric called the IOP-Score, quantifying a node's tendency to lead toward either the "Outlier" or "Inlier" class. This score is calculated as the difference between the frequency of data transitions from a node toward the "Inlier" class and those toward the "Outlier" class, normalized by the total frequency of data transitions entering the node. This normalization ensures the score accounts for the node's overall context, providing a balanced measure of its tendency to propagate toward either class. So, the IOP-Score for a generic node v is defined as:

$$\text{IOP-Score}(v) = \frac{f_i(v) - f_o(v)}{f_{in}(v)}, \tag{6.2}$$

where $f_i(v)$ is the frequency of the edge connecting node v to the "Inlier" class, $f_o(v)$ is the frequency of the edge connecting node to the "Outlier" class, and $f_{in}(v)$ is the sum of the frequencies of all edges entering node v.

- If IOP-Score$(v) = 1$, the node is fully associated with the "Inlier" class, meaning its frequency results exclusively from transitions toward the "Inlier" class. In other words, the predicate appears only in predicate lists generated by inliers traversing the model.
- If IOP-Score$(v) = -1$, the node is entirely associated with the "Outlier" class, with its frequency stemming solely from transitions toward the "Outlier" class, indicating that the predicate appears only in predicate lists generated by outliers.
- If IOP-Score$(v) = 0$, the node is considered neutral, as there is an equal frequency of transitions toward both the "Inlier" and "Outlier" classes.

In summary, an IOP-Score close to 0 indicates that the node is non-discriminative, while values near 1 or -1 signify predicates that strongly characterize one of the two classes.

Outlined in Algorithm 2, the proposed approach is presented in pseudocode to enhance clarity and understanding.

Algorithm 2: iForest as a graph for DPG-based Explanation.

Input: Trained iForest model IF, Dataset D, maximum depth of trees $dmax$
Output: iForest DPG

```
Initialize empty graph G;
foreach base learner iTree (iT) in IF do
    Extract split rules defining predicates (f, σ, v);
    foreach training sample s traversing iT do
        Record satisfied predicate lists;
        if s classified as outlier then
            Label list as "Outlier";
        else
            Label list as "Inlier";
Remove paths of predicates exceeding tree dmax for outliers;
Transform predicate lists to pairs (f, σ);
foreach predicate pair (p_i, p_j) appearing consecutively do
    Create directed edge (p_i → p_j) with frequency weight;
Apply class-based frequency weighting using ComputeFrequencyWeights();
return G;
```

6.3.2 Understanding the Explanation Process

The proposed technique is designed to capture the key concept underlying iForest. In iForest, outliers are isolated more rapidly than inliers, requiring fewer splits to

Table 6.1: DPG and their Implications for Outlier/Inlier interpretation.

Component	Implication for Outlier/Inlier Detection
Node (Predicate)	Represent a decision made to identify a sample as an inlier or outlier pathway based on feature and condition.
Weighted Edge	Indicates how frequently a decision path is used. Thicker edges leading to outliers highlight important anomaly detection features.
Node (Terminal)	Base on classified samples as inliers or outliers, helping identify critical predicates for outliers separation.
IOP-Score	Predicates with negative IOP-Scores correspond to features that play a major role in isolating outliers, while positive values indicate features that help define inlier boundaries.

separate them from the rest of the dataset. Although selecting features and associated values at each split is random, outliers differ from inliers for certain features. These key features play a role in the splits that lead to the isolation of outliers. Our XAI method's purpose is to identify the features that differentiate outliers from inliers and understand their role in IF's decision-making process. Representing the process as a graph enables visualization of predicate sequences leading to each class, highlighting the typical paths of outliers. By incorporating information about the sign of the predicates, the method enables the interpretation of the direction of the constraints imposed by the model—that is, whether a feature contributes to the isolation of outliers by surpassing a certain threshold. Moreover, using the IOP-Score—calculated for each node of the graph—quantifies the relative contribution of features in distinguishing between the two classes. A low value of this metric indicates that the corresponding predicate is essential for isolating outliers, emphasizing that the outlier nature of observations depends on specific features. This aspect underscores the importance of correctly interpreting these features within the context of the application domain and the need to consider potential data errors that may affect the identification of outliers. Furthermore, the weight of the edges connecting the nodes—proportional to frequencies—indicates whether the predicates are immediately effective at distinguishing outliers, such as when outliers are easily separated along a feature, or whether they contribute indirectly by forming decision paths that require additional splits to isolate an outlier. By combining the graph structure with IOP-Score, the proposed technique provides a global and interpretable explanation of the model. It illustrates not only which features are used but also how and with what frequency they contribute to isolating outliers.

Table 6.1 summarizes how to interpret the DPG structure to understand its implications for outlier and inlier classification.

6.4 Experiments

This section demonstrates the novelty and contributions of our DPG-based approach to explaining the iForest. We utilized a synthetic dataset to construct challenging outliers featuring multiple attributes across various scales. Additionally, we employed a benchmark dataset to facilitate a fair comparison with other techniques. This benchmark dataset was also used in the original iForest study. We conducted a comprehensive analysis, utilizing both visualizations and interpretations provided by our method.

Our implementation was developed in Python, leveraging a suite of libraries to facilitate outliers detection, visualization, and data processing. The scikit-learn library [205] was utilized for the implementation of the iForest algorithm, while Graphviz enabled the visualization of the DPG[1] , enhancing the interpretability of the decision-making process. To promote reproducibility and facilitate further research, the complete source code is publicly available on GitHub[2].

6.4.1 Synthetic datasets

To analyze our XAI methods, we generated two synthetic datasets. Each dataset contains 200 data points characterized by six numerical features (denoted as F_i, where i ranges from 1 to 6), all forming a single-cluster distribution. We introduced outliers by randomly selecting samples and modifying specific feature values according to predefined rules. Each outlier is generated by altering two and four feature values from a randomly selected sample among the available ones. Each alteration is performed by rescaling the original value by a factor of 4 or 5 times the standard deviation of that feature computed over the entire dataset. The resulting dataset exhibits clearly defined outliers, distinct feature variations, and a balanced level of complexity, making it well-suited for assessing explanation techniques in outliers detection. We trained an iForest model with 200 trees for each study case to identify outliers. Since our focus is on XAI—where the primary objective is to explain the model's decisions rather than optimize predictive accuracy—the exact number of trees is not relevant to our scope. Therefore, we chose 200 trees to ensure robust and stable predictions.

Synthetic dataset with one outlier.

The first dataset was generated by modifying four features of one sample, as reported in the Table 6.2, thereby producing a single outlier among 200 samples. In Figure 6.2, we present a pair plot of the first synthetic dataset, where we can observe that the single outline stands apart from the clustered inliers.

The modified sample was correctly identified as an outlier by the iForest model. Then, applying our technique, we obtained iForest DPG, as shown in Figure 6.3, where the classes outliers and inliers are distinguished by different colors. For each node, the IOP-Score was computed and represented by its color—these scores are summarized in Table 6.3.

[1]Implementation available at: https://github.com/LeonardoArrighi/DPG
[2]Implementation available at: https://github.com/Math0097/DPG-\gls{iforest}

Table 6.2: Sample 0 is the outlier in the first synthetic dataset. The table presents both the initial and final values of the modified features for this sample, along with the specific modifications applied to introduce the outlier.

Outliers	Feature	Initial Value	Final Value	Alteration
Sample 0	F_0	-2.12	2.29	$+4.41$
	F_3	4.05	-0.76	-4.81
	F_4	-6.01	-0.93	$+5.08$
	F_5	-7.21	-1.88	$+5.33$

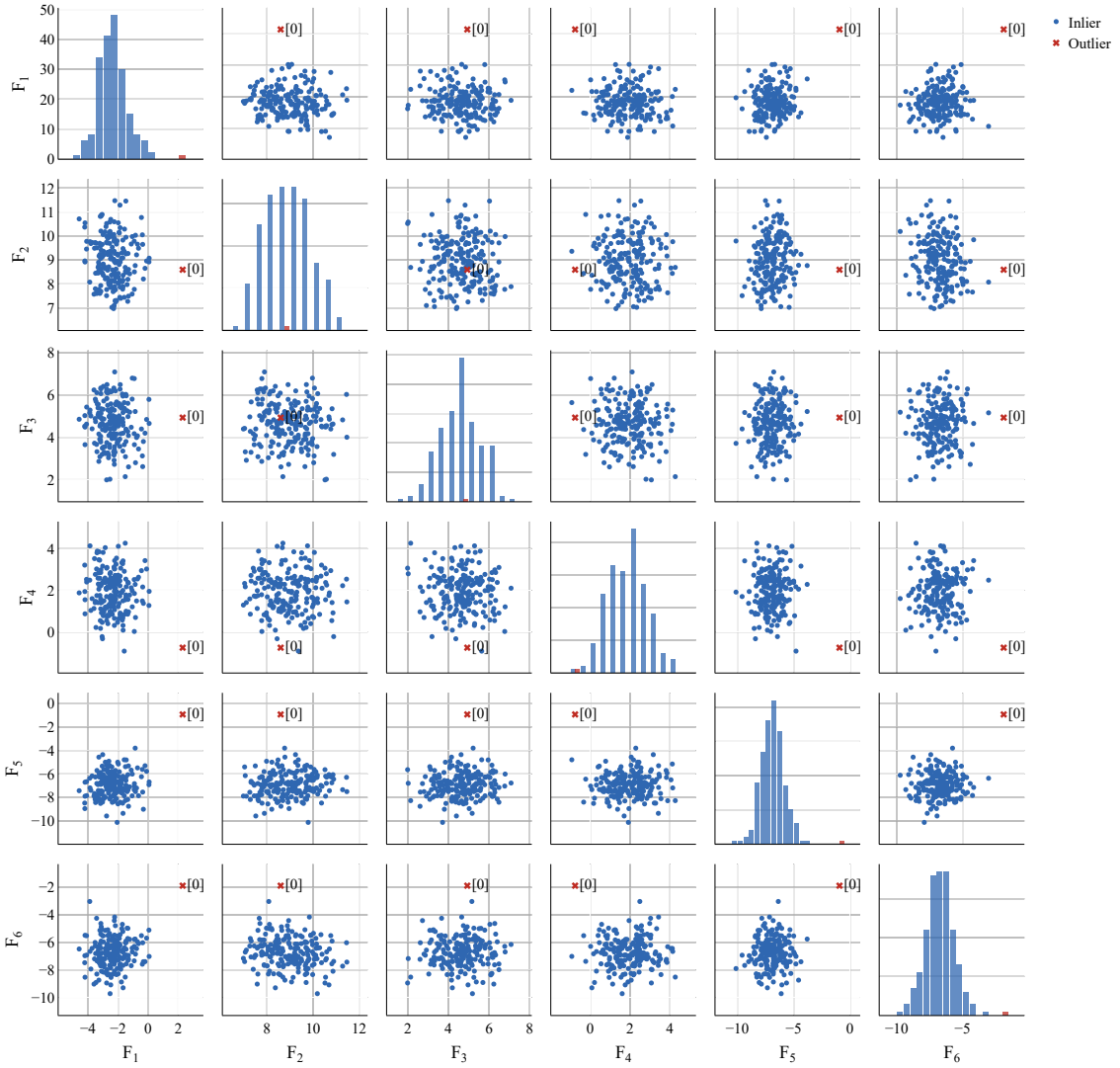


Figure 6.2: Two-dimensional representation of the first synthetic dataset. The dataset comprises 200 samples with six numerical features and one outlier.

By examining an in-depth view of the iForest model's internal process, we can observe that some nodes exhibit IOP-Score values below 0, indicating an associ-

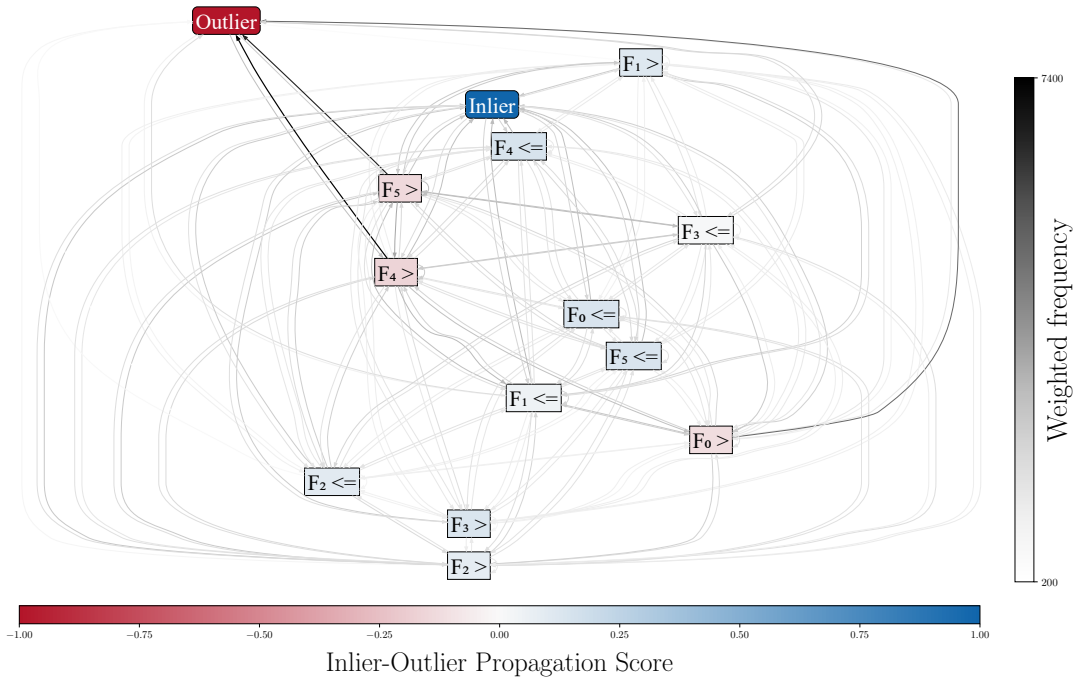


Figure 6.3: Global representation of the iForest model as a DPG for the first synthetic dataset. $F_4 >$, $F_5 >$, and $F_0 >$ have low IOP-Score, and the edges connecting them with "Outlier" class are thicker, highlighting the model's use of these features to isolate the outlier present in the dataset.

Table 6.3: IOP-Score values assigned to each predicate (node) extracted from the DPG graph of the iForest model for the first synthetic dataset. The scores quantify a node's propensity to distinguish data toward the inliers (positive values) or outliers (negative values) class.

Predicate	IOP-Score
$F_4 <=$	0.1427
$F_0 <=$	0.1406
$F_5 <=$	0.1336
$F_3 >$	0.1304
$F_1 >$	0.1129
$F_2 <=$	0.0985
$F_2 >$	0.0807
$F_1 <=$	0.0426
$F_3 <=$	0.0091
$F_0 >$	−0.1202
$F_5 >$	−0.1362
$F_4 >$	−0.1580

ation with the "Outliers" class. The nodes with the lowest scores contain particularly meaningful predicates—namely, $F_4 >$, $F_5 >$, and $F_0 >$—which correspond to the features altered to create the outliers. The $>$ sign indicates that, for the

Table 6.4: The first column lists the outliers in the second synthetic dataset. The table shows the initial and final values of the modified features for these samples, along with the specific modifications applied to introduce the outliers.

Outliers	Feature	Initial Value	Final Value	Alteration
Sample 0	F_0	-1.86	1.67	$+3.53$
	F_1	8.92	12.84	$+3.93$
Sample 1	F_0	-2.19	1.34	$+3.53$
	F_2	4.74	0.78	-3.95
Sample 2	F_0	-2.12	2.29	$+4.41$
	F_3	4.05	-0.76	-4.81
	F_5	-7.21	-1.88	$+5.33$
	F_4	-6.01	-0.93	$+5.08$
Sample 3	F_1	9.21	13.13	$+3.93$
	F_3	0.95	-2.90	-3.84

outlier, these feature values exceed those of inliers, a fact further supported by the Figure 6.2. Moreover, the edges connecting these nodes to the "Outlier" class are thicker, reflecting higher weighted frequencies; this suggests that the model consistently employs splits based on these predicates as final decision points to isolate outliers. In contrast, nodes involving predicates on F_3, despite it being one of the modified features, do not have low IOP-Score values and are not closely associated with the "Outlier" class. This indicates that F_3 does not consistently separate the irregular sample from inliers, though it does contribute to the isolation process on several occasions. Finally, the remaining nodes with IOP-Score values above 0 are predominantly involved in splits that classify points as inliers.

Synthetic dataset with four outliers.

The second dataset is created by modifying four randomly selected samples according to the previously described rule, as detailed in Table 6.4. Figure 6.4 presents an overview of the entire dataset, highlighting four outliers. Unlike the previous dataset, this one is more complex because each outlier is generated by modifying different features. As a result, each outlier can be individually distinguished by a specific set of features, meaning that no single split can separate all outliers from the inliers.

The trained iForest model successfully distinguished the modified samples as outliers. Moreover, features exhibiting consistent directional changes, such as increases in F_0 and F_1 or decreases in F_3, are more readily distinguishable than others. Similarly, as for the previous dataset, we applied our technique to explain the iForest process. The model is converted into the DPG shown in Figure 6.5, where the classes "Outlier" and "Inlier" are distinguished by different colors. For each node, the IOP-Score is computed and represented by its color—these scores are summarized in

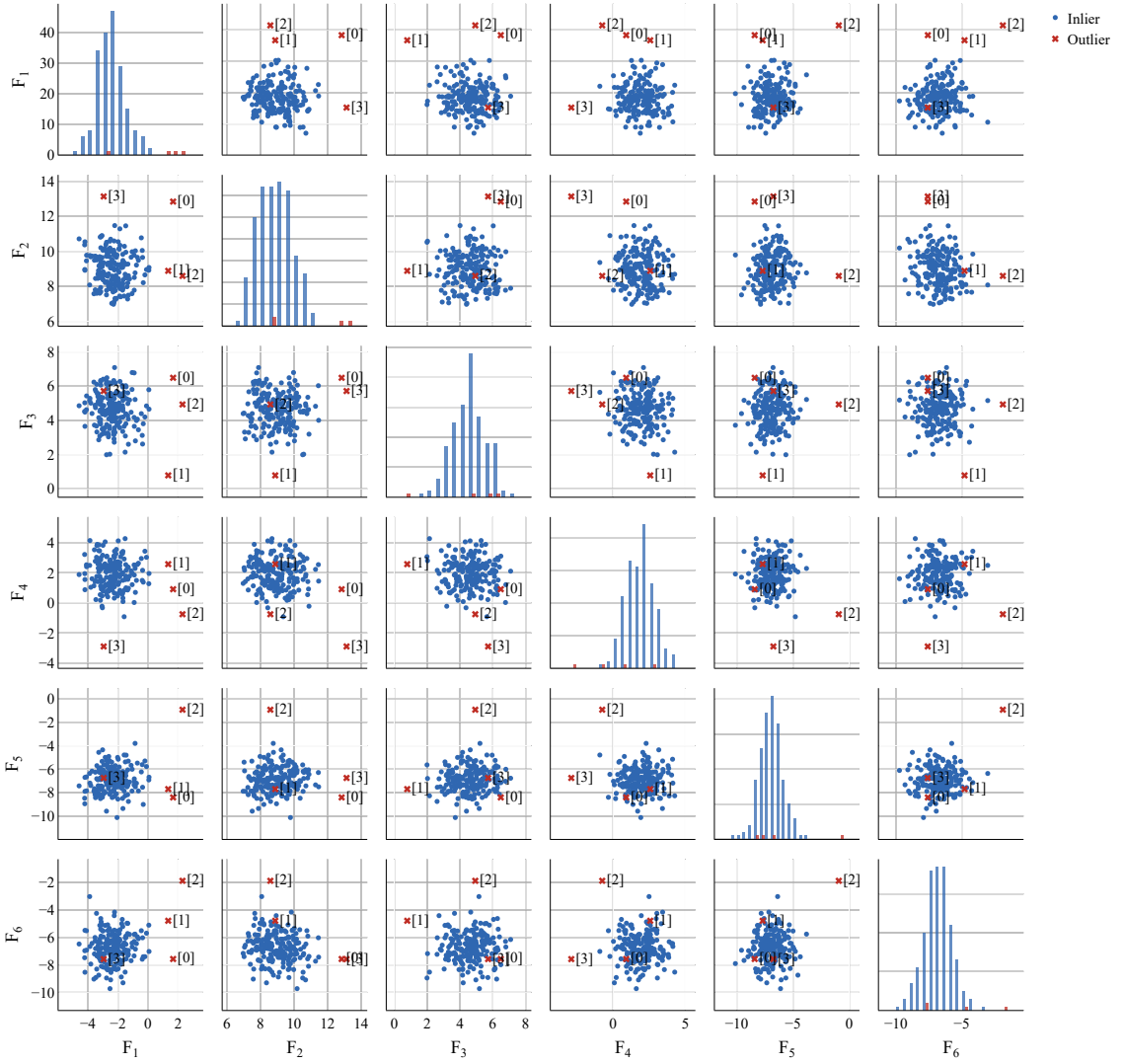


Figure 6.4: Pairplot of the second synthetic dataset. The dataset comprises 200 samples with six numerical features, and four samples have been modified by altering between two to four.

Table 6.5.

Our technique helps interpret the inner logical process of the iForest model. In this scenario, outliers are less distinct from inliers and require the combined influence of multiple features to be isolated, making the model's structure more challenging to interpret than the previous case. Nevertheless, our representation and the IOP-Score provide valuable insights. We can observe that some predicates have an IOP-Score below 0, so they are strongly connected with the "Outlier" class. In particular, the nodes with the lowest IOP-Score are $F0 >$, $F3 <=$, and $F1 >$, which comprehend the features deliberately altered to create the outliers; these predicates are critical for the model to distinguish outliers. The thicker edges connecting these nodes to the "Outlier" class further underscore their frequent use in splits that isolate irregular data points. Moreover, the directional signs in these predicates reveal how the model

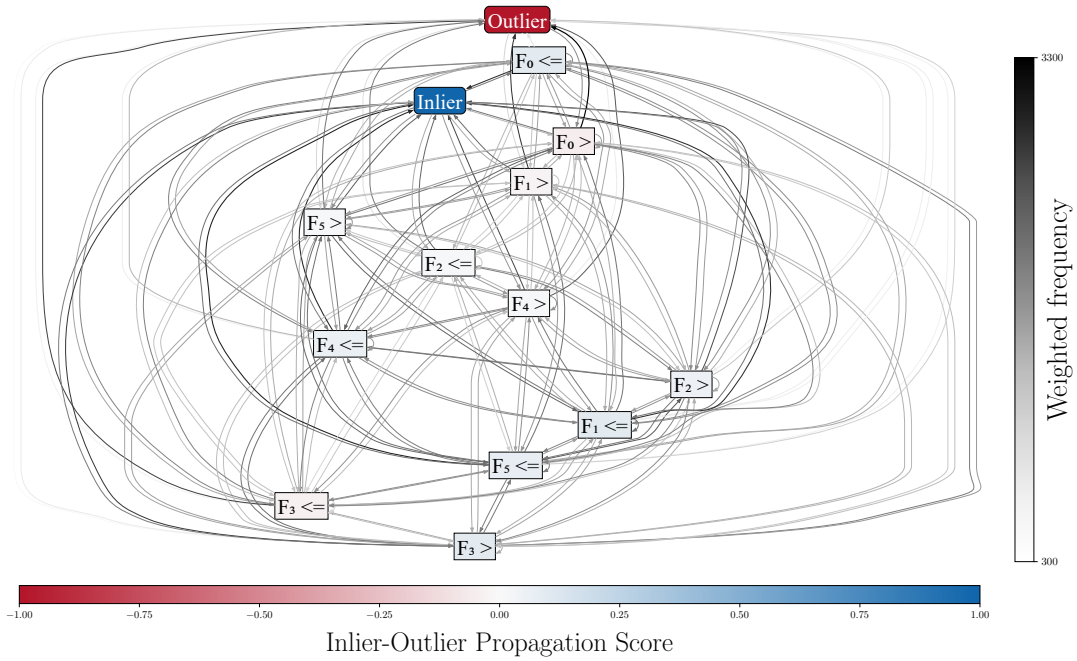


Figure 6.5: The global representation of the iForest model as a DPG for the second synthetic dataset shows that, while the contribution of specific features in identifying outliers is not very clear in this case, we can observe that $F_0 >$, $F_1 >$, and $F_3 <=$ have low IOP-Score. This confirms the model's effectiveness in isolating outliers.

Table 6.5: IOP-Score assigned to each predicate (node) extracted from the DPG graph of the iForest model for the second synthetic dataset. The scores quantify a node's propensity to channel data toward the inliers (positive values) or outliers (negative values) class.

Predicate	IOP-Score
$F_1 <=$	0.0884
$F_3 >$	0.0881
$F_0 <=$	0.0872
$F_5 <=$	0.0710
$F_4 <=$	0.0553
$F_2 >$	0.0542
$F_5 >$	0.0112
$F_2 <=$	0.0112
$F_4 >$	0.0084
$F_1 >$	−0.0257
$F_3 <=$	−0.0316
$F_0 >$	−0.0494

leverages the features—Figure 6.4 clearly shows that multiple outliers are isolated using these key splits. In addition, although F_4, F_2, and F_5 are also modified, their IOP-Scores are slightly above 0, indicating that splits involving these features do not

consistently lead to outlier isolation. Finally, the remaining nodes, with IOP-Score values above 0, are primarily involved in splits that classify points as inliers.

6.4.2 Annthyroid dataset

To evaluate the performance of our XAI methods in a real-world scenario, we used the *Annthyroid dataset*, which is widely adopted in the literature on outlier detection as a benchmark [86, 87]. The dataset represents thyroid function measurements, including hormone levels, biochemical indicators, and patient demographics. Each row corresponds to a patient sample, with multiple attributes capturing relevant physiological parameters. It consists of six numerical features (excluding the binary features) and 6916 samples. The features explored include *Age*, which provides demographic context; Thyroid-Stimulating Hormone (*TSH*), a critical regulator of thyroid function; *T3*, *TT4* (Total Thyroxine), and Free Thyroxine Index (*FTI*), which measure hormone concentrations in the blood; Thyroxine Uptake (*T4U*), which helps assess hormone-binding activity. The dataset consists of two classes: *normal* (inliers) and *anomalous* (outliers), where the latter correspond to thyroid disorders. The class distribution is highly imbalanced, with normal cases forming the majority and anomalous instances accounting for only 3.61 % of the total samples. The Annthyroid dataset is available in the UCI ML repository in the medical domain [126].

We applied our proposal to obtain an iForest model (using 200 iTrees) into a DPG and obtained results similar to the literature [87]. The explanation can be appreciated in Figure 6.6, where nodes represent predicate-based decision points while edges indicate the flow of decisions through these conditions. Thicker, darker edges correspond to frequently used decision paths, highlighting influential features, whereas lighter edges represent less significant decisions. *TSH* feature serves as a strong predicate point, with a high *TSH* value ($TSH >$) directing the flow toward the outlier node (red box). This indicates that high *TSH* levels are a significant factor in identifying thyroid anomalies with a superior limit. Similarly, a low *TSH* value ($TSH <=$) redirects the flow through additional feature-based decisions before reaching a final classification. The thin edges entering the $TSH >$ node also imply that this feature alone is usually sufficient to separate outliers from the rest of the dataset. In contrast, the $T3 >$ feature necessitates further subdivision.

The IOP-Score, in Figure 6.6, represented by the color scale at the bottom, provides further insight into how strongly each predicate affects outlier and inlier identification. Red-shaded paths and nodes indicate a high probability of leading to an outlier classification, while blue-shaded paths and nodes suggest a strong inlier association. $TSH >$ is once again revealed as a highly important factor in outliers detection. The other predicates make a slight contribution, primarily serving to delineate the boundaries of inlier behavior. More details about the obtained IOP-Score is available in Table 6.6.

Notably, as observed in Table 6.6, the node with the lowest score corresponds to the predicate $TSH >$. This is particularly significant, as its highly negative score, along with the thick edge connecting it to the "Outliers" class, suggests that the model frequently relies on this feature to isolate outliers. Similarly, the node

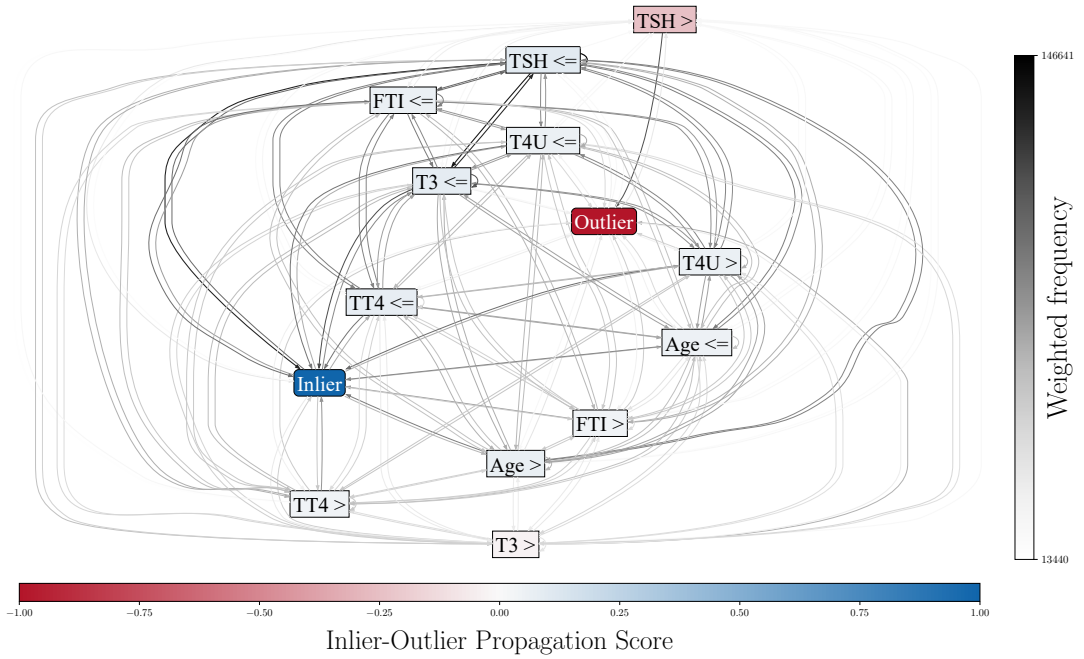


Figure 6.6: Global representation of the iForest model as a DPG structure for the Annthyroid dataset. The predicate with the lowest IOP-Score is $TSH >$, which also has the thickest incoming edge in the "Outliers" class. This indicates that the model frequently uses this feature to isolate outliers.

Table 6.6: IOP-Score assigned to each predicate (node) extracted from the DPG of the iForest model for the Annthyroid dataset. The scores quantify a node's propensity to channel data toward the "Inlier" (positive values) or "Outlier" (negative values) class.

Predicate	IOP-Score
TSH <=	0.0965
T3 <=	0.0846
TT4 <=	0.0776
Age >	0.0683
T4U <=	0.0573
FTI <=	0.0556
T4U >	0.0551
Age <=	0.0515
TT4 >	0.0354
FTI >	0.0323
T3 >	−0.0282
TSH >	−0.2429

containing the predicate $T3 >$ also has a negative score, though closer to zero, indicating that while it contributes to outlier detection, it often requires additional splits to effectively isolate outliers. Finally, the remaining nodes, with IOP-Score

values above 0, are primarily involved in splits that classify data points as inliers.

6.5 Limitations and Extensions

While the proposed approach comprehensively explains the iForest model using DPG, some limitations must be acknowledged. The transformation of iForest into a graph structure introduces additional computational complexity, mainly when dealing with high-dimensional datasets containing many trees. This complexity also leads to scalability issues, as constructing and analyzing the DPG for large-scale iForest models can be memory-intensive, necessitating optimization techniques for practical deployment. Furthermore, although DPG provides a structured representation of the model, interpreting the graph structure in highly complex datasets requires complementary visualization techniques to enhance clarity. Additionally, while existing XAI methods, such as SHAP and DIFFI provide alternative explanations for iForest, a more in-depth comparison with these techniques is necessary to establish the specific advantages and trade-offs of DPG. Another important consideration is that the method relies on predicates extracted from iForest's split rules, which may not always capture subtle feature interactions.

To address these challenges, future work will focus on optimizing graph construction techniques, improving scalability, and integrating additional interpretability metrics to enhance the usability of DPG-based explanations. The method aims to identify key features that differentiate outliers from inliers by visualizing decision paths in a graph. The incorporation of predicate signs allows for an interpretation of whether a feature contributes to outlier isolation by surpassing a threshold. Combining the graph structure with the IOP-Score enables a global understanding of the model's decision-making process, shedding light on important features and their role in detecting outliers.

6.6 Conclusion

In this work, we introduced a novel approach for explaining the iForest model using DPG. The DPG-based explanation provides a structured and interpretable representation of the outlier detection process. It offers a global perspective on the model's behavior and logic. Our approach addresses a gap in the explainability of tree-based ensemble models by extending the capabilities of traditional FI methods, such as SHAP and DIFFI, which primarily focus on local or vector-based explanations. The DPG allows for comprehensive visualization of decision paths, enabling users to interpret the isolation logic of iForest with greater clarity. Additionally, introducing the IOP-Score ensures that critical predicates contributing to outlier detection are effectively distinguished from those relevant to inliers. This paper contributes to the field of XAI by providing a transparent and interpretable method for understanding outliers detection models, offering a highly extensible approach for accurately identifying outlier behavior.

Chapter 7

End-to-End Explainability with Decision Predicate Graphs

In this Chapter, we illustrate the practicality of a complete end-to-end ML pipeline that is fully explainable through the application of the XAI techniques developed in this thesis.

The knowledge gained from this thesis project has highlighted the importance of creating a transparent pipeline. The techniques we developed, including DPG and the new method for explaining iForest, are valuable for implementing two fundamental AI-based steps that we have already addressed in this project: preprocessing and final prediction. These techniques enhance the reliability and interpretability of both processes.

However, we did not conduct the final analysis on a problem related to food quality assessment. To enhance the reliability of our results and tackle issues related to the availability of a dataset that accurately represents a real-world scenario, we chose to focus on a well-known and proven problem in a field where we recognized the need for XAI measures, specifically in finance.

Section 7.1 outlines the role of explainability in finance and reviews applications of existing techniques, setting them in relation to our proposal. In Section 7.2 we describe the proposed method and the stages of the pipeline. Section 7.3 presents the experimental study on the financial dataset, including an initial data examination. Section 7.4 reports the results and discusses the main insights obtained through explainability methods.

The content of this Chapter has been published in [11].

7.1 Introduction

AI is increasingly applied across diverse fields, with finance emerging as a prominent domain due to AI's versatility, risk-management capabilities, and its power to detect patterns in heterogeneous data sources [138, 185]. Despite their remarkable performance, many AI models operate as "black boxes", making their internal decision processes opaque. Understanding these processes in finance and other high-stakes industries is essential for ensuring trust, compliance, and reliability. XAI techniques are more used to reveal how complex models make decisions. Recent work has begun

to address this need in loan default prediction. Bracke et al. [29] applied Quantitative Input Influence with SHAP [148] to quantify feature impacts and identify key drivers of mortgage defaults. Babaei et al. [17] integrated a SHAP-based feature selection into RF models [31], yielding accurate and interpretable predictions for small and medium enterprises lending. More recently, Li and Wu [135] combined RF with SHAP to enhance predictive performance and highlight the most influential risk factors.

To avoid repeating the earlier discussion, we summarize the key point clearly. XAI should encompass the entire ML pipeline, starting from data preparation and continuing through to evaluation, rather than focusing solely on the final predictions. Decisions made during preprocessing—such as how to handle outliers, mitigate bias, and correct errors—can significantly impact the outcomes of the model. Therefore, it is essential to adopt a transparent approach that documents and justifies every stage of the process, from the raw data to the final predictions. This transparency is crucial for enhancing regulatory compliance and building user confidence [107, 142, 297].

In this study, we used DPG to explain both the preprocessing stage—specifically, outlier detection and removal via iForest in the DPG-based extension—and the final RF classifier. Our objective is to leverage the two DPG-based techniques to understand how each feature impacts outlier detection in the financial dataset through iForest, while also identifying potential biases in the preprocessing stage. Additionally, we analyze the RF classifier by highlighting its most informative DPG metrics. Together, these capabilities create a modular framework that employs DPG-based XAI techniques across these two critical phases of the pipeline. In a financial setting, this approach not only clarifies why a credit-scoring model flags an applicant as high risk but also traces how upstream cleaning steps influence the final outcome. This transparency enhances trust, especially in environments governed by regulatory frameworks such as the European Union's GDPR, which mandates the right to explanation in automated decision-making [226]. Furthermore, DPG provides interpretable, logic-based representations that compliance officers and domain experts can readily review without deep technical expertise, ensuring our framework delivers true end-to-end explainability. The main contributions of this proposal are:

- Outlier Detection and Explanation: We use iForest to detect and remove outliers, then apply DPG-extended to iForest to produce a global, feature-level explanation of their detection, clarifying each feature's impact and uncovering potential biases during data cleaning.
- Model Training and Explanation: We train an RF model on the cleaned data and then apply DPG to generate a unified, ensemble-level explanation that visualises each tree's combined decision pathways.

7.2 Proposed Approach

Figure 7.1 illustrates the proposed framework's workflow, from data preprocessing to model prediction, using DPG in our financial case study context. The pipeline begins with a raw dataset, which is first processed by an outlier detection algorithm, specifically, iForest. As seen, iForest is an unsupervised algorithm that recursively

partitions the feature space, exploiting the fact that outliers, being both rare and distinct, require fewer random splits to be isolated and thus identified. During this preprocessing phase, a DPG-based explanation for iForest is generated to provide interpretable insights into the logic behind the removal of irregular data points, ensuring interpretability in the early stages of data handling. The resulting clean dataset is then passed to a model training phase, where an RF classifier is employed to perform credit risk assessment. Moreover, with the IOP-Score it is possible to compute the tendency of a predicate to lead toward the isolation of an outlier, thereby providing a nuanced view of how specific features influence the model's outlier identifications.

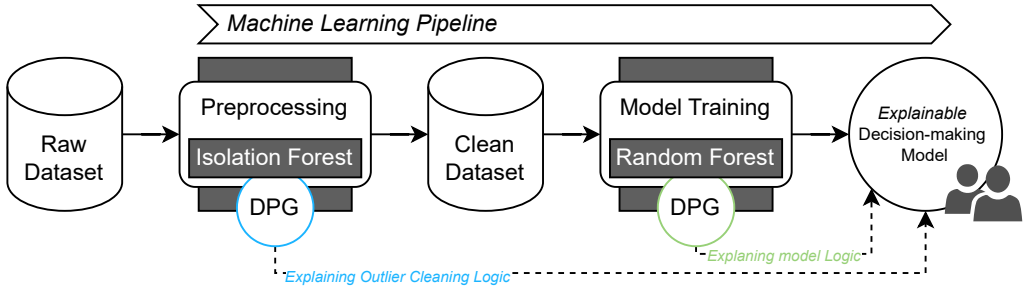


Figure 7.1: Overview of framework using DPG within an ML pipeline. DPGs are applied at both the preprocessing stage (iForest) and the model training stage (RF).

A second DPG is used to capture and explain the model's internal decision logic, highlighting the contribution of different features and their combinations to the final predictions. In fact, DPG leverages graph-theoretic metrics to provide insights into the tree-based ensemble model. It uses BC and LRC: BC identifies bottleneck nodes at critical decision points, while LRC assesses the influence of each node by measuring how its effects spread through the graph.

It is worth remembering that both techniques are represented as weighted directed graphs, where each node corresponds to a predicate and each edge is weighted by the frequency with which training samples satisfy the two predicates in sequence.

By composing predicates into graph structures, our framework captures the logical flow of decisions across the entire pipeline while also enabling graph-theoretic insights. The aim of the proposed approach is to addresses a key limitation of conventional XAI methods, which typically focus only on local explanations of model predictions, by offering a global and stepwise explanation that spans the entire ML pipeline.

7.3 Methods

Due to the privacy and accessibility constraints commonly associated with real-world financial datasets, we synthesised a dataset that simulates realistic consumer loan application scenarios while preserving reproducibility. The dataset contains $n = 400$ samples, each characterised by four features and a binary `ApprovedLoan` label. The

generation procedure was designed to reflect plausible demographic and financial distributions observed in credit risk assessment contexts[1]:

- Income (`Income`): Modelled using a log-normal distribution with a log-mean corresponding to \$60 000, and a moderate dispersion ($\sigma = 0.4$). This choice reflects the heavy-tailed nature of income in the population. Generated values were clipped to lie within the \$20 000–\$150 000 range to avoid extreme outliers.
- Credit Score (`CreditScore`): Sampled from a Gaussian distribution centered at 680 (standard deviation of 70) and truncated between the typical credit scoring bounds of 300 and 850. This ensures a realistic spread of creditworthiness scores.
- Marital Status (`MaritalStatus`): A categorical variable drawn from a discrete distribution with the following probabilities: Single (0.4), Married (0.5), and Divorced (0.1). These proportions are consistent with general demographic statistics in adult populations.
- Number of Dependent Children (`NumChildren`): Modelled using a Poisson distribution, with the expected number of children conditioned on marital status—lower for single applicants ($\lambda = 0.5$), higher for married ($\lambda = 1.2$), and moderate for divorced ($\lambda = 1.0$). Values were clipped to the $[0, 5]$ range to reflect typical household sizes.

A latent *loan approval score* was computed as a weighted sum of the input variables:

$$\begin{aligned}\text{Score} = 0.00001\,\text{Income} + 0.005\,\text{CreditScore} + 0.3\mathbb{1}_{\text{Married}} \\ + 0.2\mathbb{1}_{\text{Divorced}} - 0.15\,\text{NumChildren} + \epsilon\end{aligned}$$

where $\epsilon \sim \mathcal{N}(0, 0.1)$ is an additive noise term. The approval probability for each application was then computed using the logistic sigmoid function, centred around the number of samples to calibrate approval rates:

$$Prob_{approval} = \frac{1}{1 + \exp\left(-(Score - 4.5)\right)}$$

The binary `ApprovedLoan` label was computed from a Bernoulli distribution using $Prob_{approval}$. After synthesising the dataset, we introduced 20 controlled outliers by randomly selecting observations and increasing only their credit-score values beyond the normal range, while leaving all other features unchanged to simulate a system error. The dataset is available here.[2] The iForest model is ran with 100 trees. The RF model is trained using 100 trees. To evaluate the impact of outlier removal, we trained two RF models on different datasets: one on the full training set (including iForest-identified outliers) and one on the training set after excluding those outliers. We had split the data so that 20% of the inliers served as a common test set, and we trained both models on the remaining 80%. We then evaluated both models on the same test set to compare their performance.

[1] The features appear in brackets as they were coded in the dataset and are then used in subsequent steps to improve readability.

[2] https://github.com/LeonardoArrighi/DPG-Pipeline

7.4 Results and Discussion

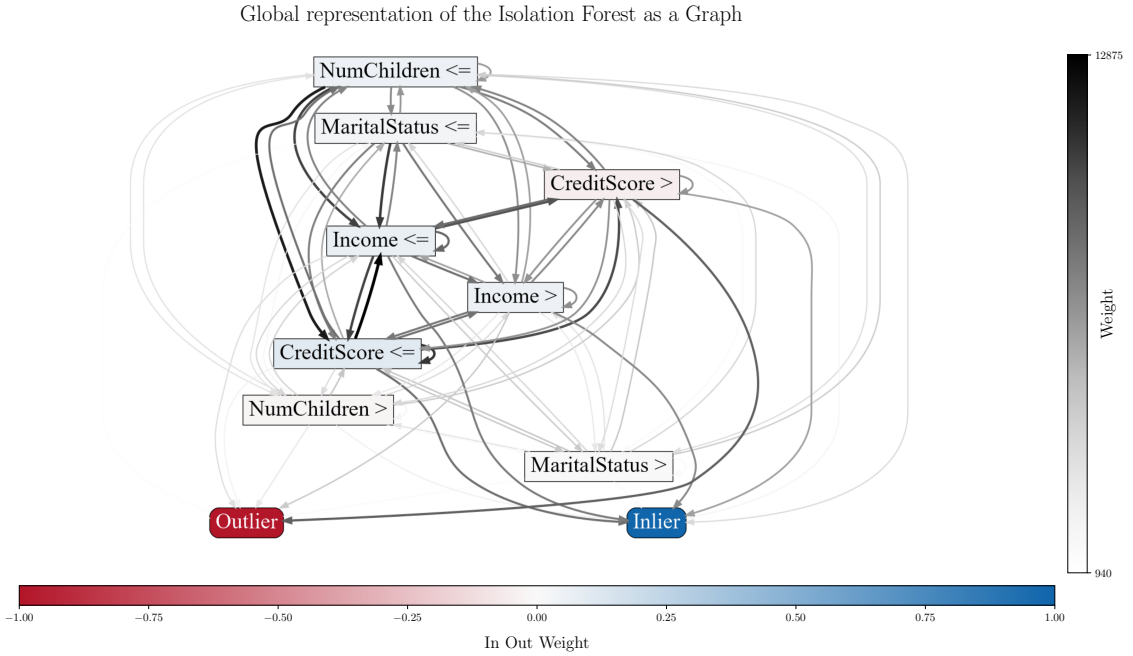


Figure 7.2: The global representation of the iForest model as DPG shows that the predicate most commonly used by the model to identify outliers is *CreditScore* >. This predicate has the lowest IOP-Score value and the highest frequency of edges leading to the "Outliers" class. This finding is expected, considering that this feature is specifically modified to generate outliers in a controlled manner.

Following the described pipeline, we first ran iForest, obtaining a set of data points identified as outliers. Applying the DPG-based technique to explain the iForest model (Figure 7.2) reveals that *CreditScore* and *Income* are the most decisive features: thicker, darker edges correspond to frequently traversed paths that highlight influential features, while thinner, lighter edges indicate less significant splits. The IOP-Score, shown on the colour scale below Figure 7.2, quantifies each predicate's impact on outlier detection. Paths and nodes shaded in red denote a high likelihood of outliers classification, whereas blue shading indicates strong association with the inlier class. Predicates such as `CreditScore >`, `NumChildren >`, and `MaritalStatus >` exhibit slightly negative IOP scores (-0.0475, -0.0204, and -0.0018, respectively), indicating a mild inclination toward directing observations to the *Outlier* class. In contrast, the remaining predicates—`MaritalStatus <=` (0.0238), `Income >` (0.0510), `NumChildren <=` (0.0545), `Income <=` (0.0673), and `CreditScore <=` (0.0958)—present positive IOP scores, suggesting they predominantly govern splits that classify observations as inliers.

We classified the dataset as *Approved* or *Not Approved* using the RF models, as described in Section 7.3. We then evaluated both models on the same test set and presented their performance in the confusion matrices of Table 7.2.

The model trained without outliers achieved a modest but significant accuracy improvement—reaching 90.79%—which confirmed that removing outliers could en-

Table 7.1: IOP-Score assigned to each predicate (node) extracted from the DPG of the iForest model for the Annthyroid dataset. The scores quantify a node's propensity to channel data toward the "Inlier" (positive values) or "Outlier" (negative values) class.

Predicate	IOP-Score
CreditScore >	−0.0475
NumChildren >	−0.0204
MaritalStatus >	−0.0018
MaritalStatus <=	0.0238
Income >	0.0510
NumChildren <=	0.0545
Income <=	0.0673
CreditScore <=	0.0958

Table 7.2: Confusion matrices that depicted the performance evaluations of the RF models with 100 base tree learners, trained on the dataset with outliers (a) and on the cleaned dataset (b).

(a) With outliers: accuracy 88.16%

	Prediction	
Ground truth	Approved	Not Approved
Approved	25	6
Not Approved	3	42

(b) Cleaned dataset: accuracy 90.79%

	Prediction	
Ground truth	Approved	Not Approved
Approved	27	4
Not Approved	3	42

Table 7.3: Top ten predicates ranked by their DPG metrics for a 100-tree RF model: (a) sorted by BC on the left; (b) sorted by LRC on the right.

(a) BC evaluation

Predicate	**BC**
`MaritalStatus <= 0.5`	0.166
`NumChildren > 0.5`	0.134
`NumChildren <= 0.5`	0.106
`MaritalStatus > 0.5`	0.090
`MaritalStatus > 1.5`	0.076
`NumChildren <= 1.5`	0.069
`MaritalStatus <= 1.5`	0.066
`CreditScore > 640.5`	0.058
`NumChildren > 1.5`	0.055
`NumChildren > 2.0`	0.043

(b) LRC evaluation

Predicate	**LRC**
`Income <= 76830.5`	9.122
`Income <= 99761.5`	9.116
`Income <= 103462.5`	9.032
`Income <= 75031.5`	8.864
`Income <= 72786.5`	8.531
`Income <= 77720.5`	8.524
`NumChildren <= 0.5`	8.504
`Income > 48467.0`	8.490
`Income > 50170.0`	8.439
`CreditScore <= 627.0`	8.188

hance predictive performance. Given RF's inherent robustness to outliers, the gain was small yet consistent.

After training the RF model, we explained it using DPG. As noted in [10], visualising complex ensembles with many tree-based learners is difficult, yet the derived metrics offer valuable insights into the model's decision mechanism.

The BC metric highlights potential bottleneck nodes, i.e., splits shared by many trees. As shown in Table 7.3a, the top features by BC are *MaritalStatus* and *NumChildren*. This result is expected, as both variables take on only a handful of discrete values, making them natural split points across the ensemble. In particular, the model relies mainly on splits at low values of these variables, underscoring that having children and being single strongly influence decisions. The third most central feature is *CreditScore*, whose split occurs near the data median; this threshold almost bisects the dataset into the two classes.

The LRC metric identifies the predicates that drive the RF's decision flow most critically. Table 7.3b ranks high-income splits on *Income* at the top, indicating that many model decisions depend on this feature. Unlike BC, *MaritalStatus* and *NumChildren* do not appear among the top LRC predicates—except for `NumChildren <= 0.5`, which still emerges as a key boundary—reflecting their lesser direct impact on the final classification.

7.5 Conclusion

This work introduced a framework for end-to-end explainability in ML pipelines, tailored to the financial domain. By integrating DPG with both outlier detection via iForest and classification using RF, we demonstrated how interpretable logic-based representations can clarify the influence of data preprocessing decisions and model predictions alike.

Chapter 8

Conclusion

In this final Chapter, we present the concluding thoughts of the thesis. We summarize the contributions of the doctoral work in Section 8.1, outline its limitations, and highlight open research directions that offer opportunities for future development and complementary solutions in Section 8.2.

8.1 Contributions

In this thesis, we recognized the importance of explainability in AI applications to make the decision-making of algorithms understandable in high-risk settings. This need reflects growing expectations for reliability and transparency, as shown by European regulation through the GDPR and the AI Act. These qualities are becoming prerequisites for deploying AI systems, especially where safety, compliance, and public trust are involved. The food supply chain is one area where AI is increasingly adopted. In this context, quality and safety decisions must be accurate and interpretable end-to-end, from data preparation to model output.

As the first step of this thesis, we conducted an analysis of the existing literature on XAI applications in one of the most critical stages of the food supply chain: food quality analysis. Our structured review revealed a fragmented body of literature and a distinct gap between food engineering practices and XAI methods. To address this gap, we proposed a domain-aware taxonomy that categorizes applications based on food quality tasks and data modalities. This taxonomy provides a common language for food scientists and data scientists. Our synthesis clarifies where explainability is currently adding value and where it is still lacking, establishing a foundation for ongoing progress in the field.

The analysis revealed that tabular data encoding physicochemical measurements are rarely utilized in XAI and food quality analysis. However, this type of data is relatively easy to collect in large quantities and can be effectively processed using tree-based ensemble models, such as RF or XGBoost. These models are particularly valuable due to their robustness and low computational cost during the training phase. Therefore, tabular data combined with tree-based ensemble models provides a solid foundation for developing a reliable pipeline in the field of food quality analysis.

From the literature review, we also found that techniques that offer a global ex-

planation for tree-based ensemble methods are few and nonspecific, although this type of explanation offers a complete overview of the model's decision-making process, thus indicating potential biases, errors in the learning process, or potentially useful insights. Building on these premises, we introduced the DPG, a novel, global, model-specific technique for tree-based ensembles. DPG converts an ensemble into a directed, weighted graph of predicates, enabling graph-theoretic analyses such as centrality and community detection. DPG overcomes a core limitation of local methods by exposing, at the ensemble level, which feature threshold combinations drive decisions. We demonstrated these benefits empirically and outlined possibilities for scaling and extension.

We then applied the method in a real-world scenario: fruit quality analysis. Using tabular physicochemical features, RF achieved strong performance on ripeness classification across three different fruit datasets (carambola, pitaya, and papaya), while DPG (global) and BELLATREX (local) together delivered "glocal" interpretability. In addition to performance, the explanations showed the method's validity, highlighting, for instance, the importance of pH and related factors in pitaya maturation. This establishes a connection between the model's logic and the mechanisms in the domain, as well as operational decisions in the supply chain.

We discussed that end-to-end AI explainability must also include transparent preprocessing. To achieve this, we extended DPG to iForest algorithm, providing the first global explanation—using DPG with an IOP-Score—of how outliers are identified, and distinguishing predicates that drive isolation from those that characterize inliers. This approach addresses a common oversight: while models are often explained, the data-cleaning processes that shape them remain opaque. Our extension makes the early stages of the pipeline as interpretable as the final predictor.

Finally, because suitable food-quality datasets were not accessible during this work, we validated the approach in another high-stakes setting: finance. We assembled a fully explainable pipeline, applying DPG both to iForest-based outlier cleansing and to RF credit scoring. The case study shows how the XAI techniques clarify cleaning decisions and classifications, supporting compliance, model-risk management, and stakeholder trust. As food datasets become available, the pipeline is ready to be transferred back to food engineering with appropriate adaptation.

8.2 Limitations and Open Research Directions

In this study, we found that DPG is an effective method for providing a global interpretation of tree-based ensemble models. However, we also identified several limitations that could guide future research. The current implementation of DPG can be computationally intensive, especially when dealing with large forests and extensive datasets. As the size of the ensembles increases and the complexity of the predictions grows, both the construction of the graphical representations and subsequent analyses become more complicated. Therefore, enhancing scalability is a top priority. Optimizing the code and developing a comprehensive library would make it possible to implement DPG on an industrial scale.

In situations like the ones described, another practical limitation often occurs when models are trained on high-dimensional datasets. In this scenarios, the graph

becomes overly dense, which limits its usability. Our focus will be on creating an interaction-based visualization and arranging the graph nodes differently to enhance clarity and readability. The objective is to develop a tool that remains navigable and informative, even when dealing with large feature spaces and numerous predicates.

Another important point to consider is that the DPG was developed and validated specifically for classification tasks. Extending its application to regression would broaden its scope significantly, allowing for end-to-end explanations in scenarios where continuous quality attributes need to be predicted.

The DPG structure is a complex object that can be used to explore deeper metrics. In addition to the centrality and community measures already examined, we plan to develop new metrics that can enhance the model's explanatory power and provide further insights.

The thesis presented specific use cases to support the discussion and validate the methods: classification of fruit ripeness using tabular physicochemical data and an end-to-end pipeline for financial risk assessment. With the knowledge gained from the literature review, we plan to apply the developed techniques to other case studies, such as novel food products, as well as to other food quality analysis tasks, including traceability and food safety.

Due to the unavailability of suitable food quality datasets during this work, the end-to-end pipeline was completed and validated in the financial domain. The next immediate step is to transfer the entire pipeline—DPG for preprocessing with iForest and DPG for the predictive model—from the financial domain to the food domain. This process involves reproducing the outlier cleaning explanations and global classifier-level explanations using real production data, as well as stress-testing the approach against the regulatory and operational constraints typically found in food processing plants.

To facilitate this, the next step is to implement the pipeline within a certified food supply chain. We plan to deploy the pipeline in compliance with established standards, ensuring that DPG explanations are integrated with existing documentation, traceability systems, and workflows for managing non-compliance. This will require alignment with new regulations promoted by the European Commission, transforming the explanations into evidence that is suitable for auditors and regulators.

Finally, to encourage adoption, we plan to release reference implementations along with benchmarks, evaluations using various XAI metrics, and clear defaults for large-scale environments.

List of Figures

List of Tables

Bibliography

[1] M. Abd Rahman and B. Ahmad Hafiz. Genetic improvement of fruit quality traits in starfruit (averrhoa carambola) hybrids. In *Acta Horticulturae*, pages 259–264. International Society for Horticultural Science (ISHS), Leuven, Belgium, 2013. ISBN 2406-6168. doi: 10.17660/ActaHortic.2013.1012.30.

[2] A. Adadi and M. Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6:52138–52160, 2018. ISSN 2169-3536. doi: 10.1109/ACCESS.2018.2870052.

[3] M. S. Ahmed, M. T. Tazwar, H. Khan, S. Roy, J. Iqbal, M. G. Rabiul Alam, M. R. Hassan, and M. M. Hassan. Yield response of different rice ecotypes to meteorological, agro-chemical, and soil physiographic factors for interpretable precision agriculture using extreme gradient boosting and support vector regression. *Complexity*, 2022:1–20, 2022. ISSN 1099-0526, 1076-2787. doi: 10.1155/2022/5305353.

[4] S. Ahmed, M. B. Hasan, T. Ahmed, M. R. K. Sony, and M. H. Kabir. Less is more: Lighter and faster deep neural architecture for tomato leaf disease classification. *IEEE Access*, 10:68868–68884, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3187203.

[5] T. Ahmed, M. B. Hasan, S. Ahmed, and M. H. Kabir. ExE-net: Explainable ensemble network for potato leaf disease classification. In *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 335–339, 2024. doi: 10.1109/CCECE59415.2024.10667205. ISSN: 2576-7046.

[6] M. Anker, C. Borsum, Y. Zhang, Y. Zhang, and C. Krupitzer. Using a machine learning regression approach to predict the aroma partitioning in dairy matrices. *Processes*, 12(2):266, 2024. ISSN 2227-9717. doi: 10.3390/pr12020266.

[7] A. Arcudi, D. Frizzo, C. Masiero, and G. A. Susto. Enhancing interpretability and generalizability in extended isolation forests. *Engineering Applications of Artificial Intelligence*, 138:109409, 2024. ISSN 0952-1976. doi: 10.1016/j.engappai.2024.109409.

[8] M. Aria, C. Cuccurullo, and A. Gnasso. A comparison among interpretative proposals for random forests. *Machine Learning with Applications*, 6:100094, 2021. ISSN 2666-8270. doi: 10.1016/j.mlwa.2021.100094.

[9] L. Arrighi, S. Barbon Junior, F. A. Pellegrino, M. Simonato, and M. Zullich. Explainable Automated Anomaly Recognition in Failure Analysis: is Deep Learning Doing it Correctly? In *Explainable Artificial Intelligence*, Communications in Computer and Information Science, pages 420–432. Springer Nature Switzerland, 2023. ISBN 978-3-031-44067-0. doi: 10.1007/978-3-031-44067-0_22.

[10] L. Arrighi, L. Pennella, G. Marques Tavares, and S. Barbon Junior. Decision predicate graphs: Enhancing interpretability in tree ensembles. In *World Conference on Explainable Artificial Intelligence*, pages 311–332. Springer Nature Switzerland, 2024. ISBN 978-3-031-63797-1. doi: 10.1007/978-3-031-63797-1_16.

[11] L. Arrighi, M. Camilo Da Silva, and S. Barbon Junior. End-to-End Explainability of Machine Learning Pipelines with Decision Predicate Graphs: A Financial Scenario Case Study. In *Ital-IA 2025 CINI National Conference on Artificial Intelligence*, 2025.

[12] L. Arrighi, I. A. de Moraes, M. Zullich, M. Simonato, D. F. Barbin, and S. Barbon Junior. Explainable artificial intelligence techniques for interpretation of food datasets: A review, 2025. URL https://arxiv.org/abs/2504.10527. *(Under Review)*.

[13] L. Arrighi, C. Giaccari, I. A. de Moraes, D. F. Barbin, and S. Barbon Junior. Enhancing Transparency in the Fruit Supply Chain Using eXplainable Artificial Intelligence, 2025. *(Under Review)*.

[14] L. Arrighi, I. A. de Moraes, M. Simonato, and S. Barbon Junior. Discriminating Short-Term Moisture Changes in Stuffed Pasta Using Deep Computer Vision. In E. Rodolà, F. Galasso, and I. Masi, editors, *Image Analysis and Processing - ICIAP 2025 Workshops*, pages 489–496, Cham, 2026. Springer Nature Switzerland. ISBN 978-3-032-11381-8. doi: 10.1007/978-3-032-11381-8_40.

[15] B. Ashoka S, M. Pramodha, A. Y. Muaad, R. Nyange, A. Anusha, N. Shilpa G, and C. Chola. Explainable AI based framework for banana disease detection. In *2024 5th International Conference on Innovative Trends in Information Technology (ICITIIT)*, pages 1–6, 2024. doi: 10.1109/ICITIIT61487.2024.10580364.

[16] E. Ayan. Genetic algorithm-based hyperparameter optimization for convolutional neural networks in the classification of crop pests. *Arabian Journal for Science and Engineering*, 49(3):3079–3093, 2024. ISSN 2191-4281. doi: 10.1007/s13369-023-07916-4.

[17] G. Babaei, P. Giudici, and E. Raffinetti. Explainable FinTech lending. *Journal of Economics and Business*, 125-126:106126, 2023. ISSN 0148-6195.

[18] S. Bach, A. Binder, G. Montavon, F. Klauschen, K.-R. Müller, and W. Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise rel-

evance propagation. *PLOS ONE*, 10(7):e0130140, 2015. ISSN 1932-6203. doi: 10.1371/journal.pone.0130140.

[19] A. Badshah, B. Yousef Alkazemi, F. Din, K. Z. Zamli, and M. Haris. Crop classification and yield prediction using robust machine learning models for agricultural sustainability. *IEEE Access*, 12:162799–162813, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3486653.

[20] X. Ban, P. Liu, L. Xu, and J. Zhao. A lightweight model based on yolov8n in wheat spike detection. In *2023 11th International Conference on Agro-Geoinformatics (Agro-Geoinformatics)*, pages 1–6, 2023. doi: 10.1109/Agro-Geoinformatics59224.2023.10233526.

[21] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020. ISSN 1566-2535. doi: 10.1016/j.inffus.2019.12.012.

[22] S. Bengamra, E. Zagrouba, and A. Bigand. Explainable AI for deep learning based potato leaf disease detection. In *2023 IEEE International Conference on Fuzzy Systems (FUZZ)*, pages 1–6, 2023. doi: 10.1109/FUZZ52849.2023.10309803. ISSN: 1558-4739.

[23] L. Benos, A. C. Tagarakis, G. Dolias, R. Berruto, D. Kateris, and D. Bochtis. Machine learning in agriculture: A comprehensive updated review. *Sensors*, 21(11):3758, 2021. ISSN 1424-8220. doi: 10.3390/s21113758.

[24] S. A. Bhat, I. Hussain, and N.-F. Huang. Soil suitability classification for crop selection in precision agriculture using GBRT-based hybrid DNN surrogate models. *Ecological Informatics*, 75:102109, 2023. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2023.102109.

[25] C. Bi, S. Xu, N. Hu, S. Zhang, Z. Zhu, and H. Yu. Identification method of corn leaf disease based on improved mobilenetv3 model. *Agronomy*, 13(2):300, 2023. ISSN 2073-4395. doi: 10.3390/agronomy13020300.

[26] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006. ISBN 978-0-387-31073-2.

[27] S. M. Blankenship and C. Unrath. Internal ethylene levels and maturity of "delicious" and "golden delicious" apples destined for prompt consumption. *Journal of the American Society for Horticultural Science*, 113(1):88–91, 1988. ISSN 0003-1062, 2327-9788. doi: 10.21273/jashs.113.1.88.

[28] V. N. Borroni, S. Fargion, A. Mazzocchi, M. Giachetti, A. Lanzarini, M. Dall'Asta, F. Scazzina, and C. Agostoni. Food quality, effects on health and sustainability today: a model case report. *International Journal of Food Sciences and Nutrition*, 68(1):117–120, 2017. doi: 10.1080/09637486.2016.1221385.

[29] P. Bracke, A. Datta, C. Jung, and S. Sen. Machine learning explainability in finance: an application to default risk analysis. *Bank of England working papers*, 2019.

[30] U. Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2):136–145, 2008. ISSN 0378-8733. doi: 10.1016/j.socnet.2007.11.001.

[31] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324.

[32] D. Broniatowski. Psychological foundations of explainability and interpretability in artificial intelligence, 2021.

[33] L. Butera, A. Ferrante, M. Jermini, M. Prevostini, and C. Alippi. Precise agriculture: Effective deep learning strategies to detect pest insects. *IEEE/CAA Journal of Automatica Sinica*, 9(2):246–258, 2022. ISSN 2329-9266, 2329-9274. doi: 10.1109/JAS.2021.1004317.

[34] O. Buyuktepe, C. Catal, G. Kar, Y. Bouzembrak, H. Marvin, and A. Gavai. Food fraud detection using explainable artificial intelligence. *Expert Systems*, 42(1):e13387, 2025. ISSN 1468-0394. doi: 10.1111/exsy.13387.

[35] M. Carletti, M. Terzi, and G. A. Susto. Interpretable anomaly detection with diffi: Depth-based feature importance of isolation forest. *Engineering Applications of Artificial Intelligence*, 119:105730, 2023. ISSN 0952-1976. doi: 10.1016/j.engappai.2022.105730.

[36] S. Castillo-Girones, R. Van Belleghem, N. Wouters, S. Munera, J. Blasco, and W. Saeys. Detection of subsurface bruises in plums using spectral imaging and deep learning with wavelength selection. *Postharvest Biology and Technology*, 207:112615, 2024. ISSN 0925-5214. doi: 10.1016/j.postharvbio.2023.112615.

[37] W. Castro, J. Marcato Junior, C. Polidoro, L. P. Osco, W. Gonçalves, L. Rodrigues, M. Santos, L. Jank, S. Barrios, C. Valle, R. Simeão, C. Carromeu, E. Silveira, L. A. d. C. Jorge, and E. Matsubara. Deep learning applied to phenotyping of biomass in forages with uav-based rgb imagery. *Sensors*, 20 (17):4802, 2020. ISSN 1424-8220. doi: 10.3390/s20174802.

[38] A. Çifci and I. Kırbaş. Fusion of machine learning and explainable AI for enhanced rice classification: a case study on cammeo and osmancik species. *European Food Research and Technology*, 251(1):69–86, 2025. ISSN 1438-2385. doi: 10.1007/s00217-024-04614-9.

[39] M. Ceschin, L. Arrighi, L. Longo, and S. Barbon Junior. Extending Decision Predicate Graphs for Comprehensive Explanation of Isolation Forest. In R. Guidotti, U. Schmid, and L. Longo, editors, *Explainable Artificial Intelligence*, pages 271–293. Springer Nature Switzerland, 2026. ISBN 978-3-032-08324-1. doi: 10.1007/978-3-032-08324-1_12.

[40] H. Chandra, P. M. Pawar, R. Elakkiya, P. S. Tamizharasan, R. Muthalagu, and A. Panthakkan. Explainable AI for soil fertility prediction. *IEEE Access*, 11:97866–97878, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3311827.

[41] A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian. Grad-CAM++: Improved visual explanations for deep convolutional networks. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 839–847, 2018. doi: 10.1109/WACV.2018.00097.

[42] J. Chen, D. Zhang, and Y. A. Nanehkaran. Identifying plant diseases using deep transfer learning and enhanced lightweight network. *Multimedia Tools and Applications*, 79(41):31497–31515, 2020. ISSN 1573-7721. doi: 10.1007/s11042-020-09669-w.

[43] J. Chen, D. Zhang, M. Suzauddola, and A. Zeb. Identifying crop diseases using attention embedded MobileNet-v2 model. *Applied Soft Computing*, 113: 107901, 2021. ISSN 1568-4946. doi: 10.1016/j.asoc.2021.107901.

[44] J. Chen, A. Zeb, Y. A. Nanehkaran, and D. Zhang. Stacking ensemble model of deep learning for plant disease recognition. *Journal of Ambient Intelligence and Humanized Computing*, 14(9):12359–12372, 2023. ISSN 1868-5145. doi: 10.1007/s12652-022-04334-6.

[45] L. Chen, X. Cui, and W. Li. Meta-learning for few-shot plant disease detection. *Foods*, 10(10):2441, 2021. ISSN 2304-8158. doi: 10.3390/foods10102441.

[46] R. Chen, H. Qi, Y. Liang, and M. Yang. Identification of plant leaf diseases by deep learning based on channel attention and channel pruning. *Frontiers in Plant Science*, 13, 2022. ISSN 1664-462X. doi: 10.3389/fpls.2022.1023515.

[47] T. Chen and C. Guestrin. XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939785.

[48] T. R. Chhetri, A. Hohenegger, A. Fensel, M. A. Kasali, and A. A. Adekunle. Towards improving prediction accuracy and user-level explainability using deep learning and knowledge graphs: A study on cassava disease. *Expert Systems with Applications*, 233:120955, 2023. ISSN 0957-4174. doi: 10.1016/j.eswa.2023.120955.

[49] H. Chipman, E. George, and R. McCulloch. Making sense of a forest of trees. *Proceedings of the 30th Symposium on the Interface*, 29, 1998.

[50] V. S. S. V. Chivukula, G. Anuradha, S. N. C. Dhanekula, and N. G. Kothagundla. Rice crop disease detection using explainable AI. In *2023 Global Conference on Information Technologies and Communications (GCITC)*, pages 1–8, 2023. doi: 10.1109/GCITC60406.2023.10425857.

[51] M. E. H. Chowdhury, T. Rahman, A. Khandakar, M. A. Ayari, A. U. Khan, M. S. Khan, N. Al-Emadi, M. B. I. Reaz, M. T. Islam, and S. H. M. Ali. Automatic and reliable leaf disease detection using deep learning techniques. *AgriEngineering*, 3(2):294–312, 2021. ISSN 2624-7402. doi: 10.3390/agriengineering3020020.

[52] E. Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) (Text with EEA relevance), 2016.

[53] S. Coulibaly, B. Kamsu-Foguem, D. Kamissoko, and D. Traore. Explainable deep convolutional neural networks for insect pest recognition. *Journal of Cleaner Production*, 371:133638, 2022. ISSN 0959-6526. doi: 10.1016/j.jclepro.2022.133638.

[54] Z. Cui, N. Zhang, T. Zhou, X. Zhou, H. Meng, Y. Yu, Z. Zhang, Y. Zhang, W. Wang, and Y. Liu. Conserved sites and recognition mechanisms of t1r1 and t2r14 receptors revealed by ensemble docking and molecular descriptors and fingerprints combined with machine learning. *Journal of Agricultural and Food Chemistry*, 71(14):5630–5645, 2023. ISSN 0021-8561. doi: 10.1021/acs.jafc.3c00591.

[55] M. V. da Silva Ferreira, I. A. de Moraes, R. V. L. Passos, D. F. Barbin, and J. L. Barbosa Jr. Determination of pitaya quality using portable nir spectroscopy and innovative low-cost electronic nose. *Scientia Horticulturae*, 310:111784, 2023. doi: 10.1016/j.scienta.2022.111784.

[56] M. V. da Silva Ferreira, S. Barbon Junior, V. G. Turrisi da Costa, D. F. Barbin, and J. Lucena Barbosa Jr. Deep computer vision system and explainable artificial intelligence applied for classification of dragon fruit (*Hylocereus spp.*). *Scientia Horticulturae*, 338:113605, 2024. ISSN 0304-4238. doi: 10.1016/j.scienta.2024.113605.

[57] R. Daniel-Weiner, M. I. Cardel, M. Skarlinski, A. Goscilo, C. Anderson, and G. D. Foster. Enabling informed decision making in the absence of detailed nutrition labels: A model to estimate the added sugar content of foods. *Nutrients*, 15:803, 2023. ISSN 2072-6643. doi: 10.3390/nu15040803.

[58] M. de Benito Fernández, D. L. Martínez, A. González-Briones, P. Chamoso, and E. S. Corchado. Evaluation of XAI models for interpretation of deep learning techniques' results in automated plant disease diagnosis. *Trends in Sustainable Smart Cities and Territories*, pages 417–428, 2023. doi: 10.1007/978-3-031-36957-5_36.

[59] I. A. de Moraes, L. J. P. Cruz-Tirado, and D. F. Barbin. Online measurement of carambola (*Averrhoa carambola* l.) physicochemical properties and estimation of maturity stages using a portable nir spectrometer. *Scientia Horticulturae*, 304:111263, 2022. doi: 10.1016/j.scienta.2022.111263.

[60] I. A. de Moraes, S. Barbon Junior, and D. F. Barbin. Interpretation and explanation of computer vision classification of carambola (*Averrhoa carambola L.*) according to maturity stage. *Food Research International*, 192:114836, 2024. ISSN 0963-9969. doi: 10.1016/j.foodres.2024.114836.

[61] I. A. de Moraes, L. Arrighi, S. Barbon Junior, J. E. L. Villa, R. L. Cunha, and D. F. Barbin. Explainable artificial intelligence (xAI) applied to deep computer vision of microscopy imaging and spectroscopy for assessment of oleogel stability over storage. *Journal of Food Engineering*, 394:112515, 2025. ISSN 0260-8774. doi: 10.1016/j.jfoodeng.2025.112515.

[62] K. Dedja, F. K. Nakano, K. Pliakos, and C. Vens. BELLATREX: Building explanations through a LocaLly AccuraTe rule EXtractor. *IEEE Access*, 11: 41348 – 41367, 2023. ISSN 2169-3536. doi: 10.1109/ACCESS.2023.3268866.

[63] H. Deng. Interpreting tree ensembles with inTrees. *International Journal of Data Science and Analytics*, 7(4):277–287, 2019. ISSN 2364-4168. doi: 10.1007/s41060-018-0144-8.

[64] M. Dileo, R. Olmeda, M. Pindaro, and M. Zignani. Graph machine learning for fast product development from formulation trials. *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track*, pages 303–318, 2024. doi: 10.1007/978-3-031-70378-2_19.

[65] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[66] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020.

[67] Y. Durmuş and A. F. Atasoy. Application of multivariate machine learning methods to investigate organic compound content of different pepper spices. *Food Bioscience*, 51:102216, 2023. ISSN 2212-4292. doi: 10.1016/j.fbio.2022.102216.

[68] P. Dutta, D. Jain, R. Gupta, and B. Rai. Classification of tastants: A deep learning based approach. *Molecular Informatics*, 42(12):e202300146, 2023. ISSN 1868-1751. doi: 10.1002/minf.202300146.

[69] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, and R. Ranjan. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9), 2023. doi: 10.1145/3561048.

[70] G. N. Elwirehardja, T. Suparyanto, Miftakhurrokhmat, and B. Pardamean. Determining variables associated with annual oil palm yield: An explainable

gradient boosting approach. *Procedia Computer Science*, 227:262–271, 2023. ISSN 1877-0509. doi: 10.1016/j.procs.2023.10.524.

[71] A. En-nhaili, A. Hachmoud, A. Meddaoui, and A. Jrifi. Enhancing product predictive quality control using machine learning and explainable AI. *Data and Metadata*, 4:500–500, 2025. ISSN 2953-4917. doi: 10.56294/dm2025500.

[72] E. Ennadifi, S. Laraba, D. Vincke, B. Mercatoris, and B. Gosselin. Wheat diseases classification and localization using convolutional neural networks and GradCAM visualization. *2020 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–5, 2020. doi: 10.1109/ISCV49265.2020.9204258.

[73] R. P. P. Ethiraj, Kavitha. A deep learning-based approach for early detection of disease in sugarcane plants: an explainable artificial intelligence model. *IAES International Journal of Artificial Intelligence (IJ-AI)*, 13(1):974–983, 2024. ISSN 2252-8938. doi: 10.11591/ijai.v13.i1.pp974-983.

[74] European Commission. Joint Research Centre. *AI Watch, AI standardisation landscape state of play and link to the EC proposal for an AI regulatory framework.* Publications Office, LU, 2021.

[75] M. Farras, J. R. Swann, I. Rowland, L. Rubio, I. Subirana, U. Catalan, M. J. Motilva, R. Solà, M. I. Covas, F. Blanco-Vaca, M. Fitó, and J. Mayneris-Perxachs. Impact of phenol-enriched olive oils on serum metabonome and its relationship with cardiometabolic parameters: A randomized, double-blind, cross-over, controlled trial. *Antioxidants*, 11(10):1964, 2022. ISSN 2076-3921. doi: 10.3390/antiox11101964.

[76] J. Feng and X. Xu. Deciphering plant seedlings: Enhancing classification and interpretability with vision transformers. In *2024 5th International Conference on Computer Vision, Image and Deep Learning (CVIDL)*, pages 635–640, 2024. doi: 10.1109/CVIDL62147.2024.10604151.

[77] J. Fernandes Lopes, V. G. Turrisi da Costa, D. F. Barbin, L. Cruz-Tirado, V. Baeten, and S. Barbon Junior. Deep computer vision system for cocoa classification. *Multimedia Tools and Applications*, 81(18):24987–25008, 2022. doi: 10.1007/s11042-022-13097-3.

[78] P. Filippi, B. M. Whelan, and T. F. A. Bishop. Explainable machine learning to map the impact of weather and soil on wheat yield and revenue across the eastern australian grain belt. *Agriculture*, 14(12):2318, 2024. ISSN 2077-0472. doi: 10.3390/agriculture14122318.

[79] R. A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188, 1936. ISSN 2050-1439. doi: 10.1111/j.1469-1809.1936.tb02137.x.

[80] A. M. Florio, P. Martins, M. Schiffer, T. Serra, and T. Vidal. Optimal decision diagrams for classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7577–7585, 2023. doi: 10.1609/aaai.v37i6.25920.

[81] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. ISSN 0090-5364.

[82] C. Gao, W. Guo, C. Yang, Z. Gong, J. Yue, Y. Fu, and H. Feng. A fast and lightweight detection model for wheat fusarium head blight spikes in natural environments. *Computers and Electronics in Agriculture*, 216:108484, 2024. ISSN 0168-1699. doi: 10.1016/j.compag.2023.108484.

[83] M. Gehlot and G. C. Gandhi. "EffiNet-Ts": A deep interpretable architecture using EfficientNet for plant disease detection and visualization. *Journal of Plant Diseases and Protection*, 130:413–430, 2023. ISSN 1861-3837. doi: 10.1007/s41348-023-00707-x.

[84] I. Y. Genç, R. Gürfidan, and T. Yiğit. Quality prediction of seabream *Sparus aurata* by deep learning algorithms and explainable artificial intelligence. *Food Chemistry*, 474:143150, 2025. ISSN 0308-8146. doi: 10.1016/j.foodchem.2025.143150.

[85] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin. Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. *journal of Computational and Graphical Statistics*, 24(1):44–65, 2015.

[86] M. Goldstein. Unsupervised Anomaly Detection Benchmark, 2015.

[87] M. Goldstein and S. Uchida. A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data. *PLoS ONE*, 11(4):e0152173, 2016. doi: 10.1371/journal.pone.0152173.

[88] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. ISSN 0001-0782. doi: 10.1145/3422622.

[89] F. Gossen and B. Steffen. Algebraic aggregation of random forests: towards explainability and rapid evaluation. *International Journal on Software Tools for Technology Transfer*, 25(3):1–19, 2021. ISSN 1433-2787. doi: 10.1007/s10009-021-00635-x.

[90] P. Gowri, S. Aathilakshmi, G. Sivapriya, A. Boomika, K. Ashika, and P. Aswin. Explainable AI-based model interpretability for tomato leaf disease identification. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2024. doi: 10.1109/ICCCNT61001.2024.10724346. ISSN: 2473-7674.

[91] B. Gulowaty and M. Woźniak. Extracting interpretable decision tree ensemble from random forest. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE, 2021.

[92] T. Guo, F. Pan, Z. Cui, Z. Yang, Q. Chen, L. Zhao, and H. Song. Fapd: An astringency threshold and astringency type prediction database for flavonoid compounds based on machine learning. *Journal of Agricultural and Food*

Chemistry, 71(9):4172–4183, 2023. ISSN 0021-8561. doi: 10.1021/acs.jafc.2c08822.

[93] M. Haddouchi and A. Berrado. A survey of methods and tools used for interpreting random forest. In *2019 1st International Conference on Smart Systems and Data Science (ICSSD)*, pages 1–6, 2019. doi: 10.1109/ICSSD47982.2019.9002770.

[94] A. Halabi Diaz, F. Galdames, and P. Velásquez. Accurate & simple open-sourced no-code machine learning and CDFT predictive models for the antioxidant activity of phenols. *Computational and Theoretical Chemistry*, 1239:114782, 2024. ISSN 2210-271X. doi: 10.1016/j.comptc.2024.114782.

[95] S. A. Halim-Lim, N. H. Ahmad, and N. Z. N. Hasnan. Quality and safety in the food industry. *Wiley StatsRef: Statistics Reference Online*, pages 1–7, 2022. doi: 10.1002/9781118445112.stat08389.

[96] J. Han, L. Shi, Q. Yang, K. Huang, Y. Zha, and J. Yu. Real-time detection of rice phenology through convolutional neural network using handheld camera images. *Precision Agriculture*, 22(1):154–178, 2021. ISSN 1573-1618. doi: 10.1007/s11119-020-09734-2.

[97] Y. Han, Q. Cheng, W. Wu, and Z. Huang. Dpf-nutrition: Food nutrition estimation via depth prediction and fusion. *Foods*, 12(23):4293, 2023. ISSN 2304-8158. doi: 10.3390/foods12234293.

[98] M. A. Haque, S. Marwaha, C. K. Deb, S. Nigam, and A. Arora. Recognition of diseases of maize crop using deep learning models. *Neural Computing and Applications*, 35(10):7407–7421, 2023. ISSN 1433-3058. doi: 10.1007/s00521-022-08003-9.

[99] S. Hara and K. Hayashi. Making tree ensembles interpretable: A bayesian model selection approach. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, pages 77–85. PMLR, 2018.

[100] M. Hasan, N. Vasker, and M. S. H. Khan. Real-time sorting of broiler chicken meat with robotic arm: XAI-enhanced deep learning and LIME framework for freshness detection. *Journal of Agriculture and Food Research*, 18:101372, 2024. ISSN 2666-1543. doi: 10.1016/j.jafr.2024.101372.

[101] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics. Springer, New York, NY, 2009. ISBN 978-0-387-84857-0 978-0-387-84858-7. doi: 10.1007/978-0-387-84858-7.

[102] J. Hatwell, M. M. Gaber, and R. M. A. Azad. CHIRPS: Explaining random forest classification. *Artificial Intelligence Review*, 53(8):5747–5788, 2020. ISSN 1573-7462. doi: 10.1007/s10462-020-09833-6.

[103] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. doi: 10.1109/CVPR.2016.90.

[104] M. Heino, P. Kinnunen, W. Anderson, D. K. Ray, M. J. Puma, O. Varis, S. Siebert, and M. Kummu. Increased probability of hot and dry weather extremes during the growing season threatens global crop yields. *Scientific Reports*, 13(1):3583, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-29378-2.

[105] I. Hernández, S. Gutiérrez, I. Barrio, R. Íñiguez, and J. Tardaguila. In-field disease symptom detection and localisation using explainable deep learning: Use case for downy mildew in grapevine. *Computers and Electronics in Agriculture*, 226:109478, 2024. ISSN 0168-1699. doi: 10.1016/j.compag.2024.109478.

[106] T. K. Ho. Random decision forests. In *Proceedings of 3rd International Conference on Document Analysis and Recognition*, volume 1, pages 278–282 vol.1, 1995. doi: 10.1109/ICDAR.1995.598994.

[107] A. Holzinger, A. Carrington, and H. Muller. Measuring the quality of explanations: The system causability scale (scs). *KI - Künstliche Intelligenz*, 34(2): 193–198, 2020. ISSN 1610-1987. doi: 10.1007/s13218-020-00636-z.

[108] B. Hoon Yun, H.-Y. Yu, H. Kim, S. Myoung, N. Yeo, J. Choi, H. Sook Chun, H. Kim, and S. Ahn. Geographical discrimination of asian red pepper powders using 1h nmr spectroscopy and deep learning-based convolution neural networks. *Food Chemistry*, 439:138082, 2024. ISSN 0308-8146. doi: 10.1016/j.foodchem.2023.138082.

[109] T. Hu, X. Zhang, G. Bohrer, Y. Liu, Y. Zhou, J. Martin, Y. Li, and K. Zhao. Crop yield prediction via explainable AI and interpretable machine learning: Dangers of black box models for evaluating climate change impacts on crop yield. *Agricultural and Forest Meteorology*, 336:109458, 2023. ISSN 0168-1923. doi: 10.1016/j.agrformet.2023.109458.

[110] Z. Huang, R. Wang, Y. Cao, S. Zheng, Y. Teng, F. Wang, L. Wang, and J. Du. Deep learning based soybean seed classification. *Computers and Electronics in Agriculture*, 202:107393, 2022. ISSN 0168-1699. doi: 10.1016/j.compag.2022.107393.

[111] F. Huber, A. Yushchenko, B. Stratmann, and V. Steinhage. Extreme gradient boosting for yield estimation compared with deep learning approaches. *Computers and Electronics in Agriculture*, 202:107346, 2022. ISSN 0168-1699. doi: 10.1016/j.compag.2022.107346.

[112] T. Ibrahim, K. B. Isaac, B. Francis, E. Lule, N. Hellen, H. Chongomweru, and G. Marvin. Interpretable machine learning techniques for predictive cattle behavior monitoring. In *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, pages 1219–1224, 2024. doi: 10.1109/ICSCSS60660.2024.10625182.

[113] D. Ignatov and A. Ignatov. Decision stream: Cultivating deep decision trees. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 905–912. IEEE, 2017. doi: 10.1109/ICTAI.2017.00140.

[114] S. Ittisoponpisan, C. Kaipan, S. Ruang-On, R. Thaiphan, and K. Songsri-In. Pushing the accuracy of thai food image classification with transfer learning. *Engineering Journal*, 26(10):57–71, 2022. ISSN 0125-8281. doi: 10.4186/ej.2022.26.10.57.

[115] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning: with Applications in R.* Springer Texts in Statistics. Springer US, 2021. ISBN 978-1-07-161417-4 978-1-07-161418-1. doi: 10.1007/978-1-0716-1418-1.

[116] S. Jiang, W. Min, Y. Lyu, and L. Liu. Few-shot food recognition via multi-view representation learning. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3):1–20, 2020. ISSN 1551-6857, 1551-6865. doi: 10.1145/3391624.

[117] A. Jlassi, A. Elaoud, H. Ghazouani, and W. Barhoumi. Potato leaf disease classification using transfer learning and reweighting-based training with imbalanced data. *SN Computer Science*, 5(8):987, 2024. ISSN 2661-8907. doi: 10.1007/s42979-024-03334-x.

[118] E. Jo, Y. Lee, Y. Lee, J. Baek, and J. G. Kim. Rapid identification of counterfeited beef using deep learning-aided spectroscopy: Detecting colourant and curing agent adulteration. *Food and Chemical Toxicology*, 181:114088, 2023. ISSN 0278-6915. doi: 10.1016/j.fct.2023.114088.

[119] R. John Martin, R. Mittal, V. Malik, F. Jeribi, S. Tabrez Siddiqui, M. Alamgir Hossain, and S. L. Swapna. XAI-powered smart agriculture framework for enhancing food productivity and sustainability. *IEEE Access*, 12:168412–168427, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3492973.

[120] S. Jubair, O. Tremblay-Savard, and M. Domaratzki. Gxenet: Novel fully connected neural network based approaches to incorporate gxe for predicting wheat yield. *Artificial Intelligence in Agriculture*, 8:60–76, 2023. ISSN 2589-7217. doi: 10.1016/j.aiia.2023.05.001.

[121] V. Kakani, V. H. Nguyen, B. P. Kumar, H. Kim, and V. R. Pasupuleti. A critical review on computer vision and artificial intelligence in food industry. *Journal of Agriculture and Food Research*, 2:100033, 2020. ISSN 2666-1543. doi: 10.1016/j.jafr.2020.100033.

[122] E. Kalopesa, K. Karyotis, N. Tziolas, N. Tsakiridis, N. Samarinas, and G. Zalidis. Estimation of sugar content in wine grapes via in situ vnir–swir point spectroscopy using explainable artificial intelligence techniques. *Sensors*, 23 (3):1065, 2023. ISSN 1424-8220. doi: 10.3390/s23031065.

[123] M. F. Kalyango and K. M. Ntanda. Interpretable deep learning for diagnosis of maize streak disease. In *2023 First International Conference on the Advancements of Artificial Intelligence in African Context (AAIAC)*, pages 1–6, 2023. doi: 10.1109/AAIAC60008.2023.10465315.

[124] S. Kamal, P. Sharma, P. K. Gupta, M. K. Siddiqui, A. Singh, and A. Dutt. DVTXAI: a novel deep vision transformer with an explainable AI-based framework and its application in agriculture. *The Journal of Supercomputing*, 81 (1):280, 2024. ISSN 1573-0484. doi: 10.1007/s11227-024-06494-y.

[125] N. S. Kartha, C. Gautrais, and V. Vercruyssen. Why are you weird? infusing interpretability in isolation forest for anomaly detection. *arXiv preprint arXiv:2112.06858*, 2021. doi: 10.48550/arXiv.2112.06858.

[126] M. Kelly, R. Longjohn, and K. Nottingham. UCI machine learning repository. URL http://archive.ics.uci.edu/ml.

[127] Q.-H. Kha, V.-H. Le, T. N. K. Hung, N. T. K. Nguyen, and N. Q. K. Le. Development and validation of an explainable machine learning-based prediction model for drug–food interactions from chemical structures. *Sensors*, 23 (8):3962, 2023. ISSN 1424-8220. doi: 10.3390/s23083962.

[128] S. Z. Khan, S. Dhou, and A. R. Al-Ali. Machine learning based palm farming: Harvesting and disease identification. *IEEE Access*, 12:157854–157871, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3484943.

[129] H. Kim, E. Hewett, and N. Lallu. The role of ethylene in kiwifruit softening. In *Acta Horticulturae*, pages 255–262. International Society for Horticultural Science (ISHS), Leuven, Belgium, 1999. ISBN 2406-6168. doi: 10.17660/ActaHortic.1999.498.29.

[130] I. Kollia, J. Stevenson, and S. Kollias. Ai-enabled efficient and safe food supply chain. *Electronics*, 10(11):1223, 2021. ISSN 2079-9292. doi: 10.3390/electronics10111223.

[131] O. Konda, R. A. Sharief Mohammad, S. Mishra, N. Rajeev, and A. Verma. Harvesting insights: Leveraging explainable AI to optimize farming practices. In *2024 International Conference on Advances in Computing Research on Science Engineering and Technology (ACROSET)*, pages 1–8, 2024. doi: 10.1109/ACROSET62108.2024.10743799.

[132] S. Kumar and M. Kumar. Enhancing agricultural decision-making through an explainable AI-based crop recommendation system. In *2024 International Conference on Signal Processing and Advance Research in Computing (SPARC)*, volume 1, pages 1–6, 2024. doi: 10.1109/SPARC61891.2024.10829064.

[133] N. Kundu, G. Rani, V. S. Dhaka, K. Gupta, S. C. Nayaka, E. Vocaturo, and E. Zumpano. Disease detection, severity prediction, and crop loss estimation in maizecrop using deep learning. *Artificial Intelligence in Agriculture*, 6:276–291, 2022. ISSN 2589-7217. doi: 10.1016/j.aiia.2022.11.002.

[134] B. Li, B. Liu, S. Li, and H. Liu. An improved efficientnet for rice germ integrity classification and recognition. *Agriculture*, 12(6):863, 2022. ISSN 2077-0472. doi: 10.3390/agriculture12060863.

[135] H. Li and W. Wu. Loan default predictability with explainable machine learning. *Finance Research Letters*, 60:104867, 2024. ISSN 1544-6123.

[136] J. Li, B. Zhao, J. Wu, S. Zhang, F. Wang, and C. Lv. Mbnet: A multi-branch network for detecting the appearance of korla pears. *Computers and Electronics in Agriculture*, 206:107660, 2023. ISSN 0168-1699. doi: 10.1016/j.compag.2023.107660.

[137] R. Li, J. Chen, J. Yang, and C. Wang. Explainable artificial intelligence for evaluation of liquor. In *IECON 2022 – 48th Annual Conference of the IEEE Industrial Electronics Society*, pages 1–6, 2022. doi: 10.1109/IECON49645.2022.9968447.

[138] X. Li, A. Sigov, L. Ratkin, L. A. Ivanov, and L. Li. Artificial intelligence applications in finance: a survey. *Journal of Management Analytics*, 10(4): 676–692, 2023.

[139] Y. Li, H. Zeng, M. Zhang, B. Wu, Y. Zhao, X. Yao, T. Cheng, X. Qin, and F. Wu. A county-level soybean yield prediction framework coupled with xgboost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation*, 118:103269, 2023. ISSN 1569-8432. doi: 10.1016/j.jag.2023.103269.

[140] H. Liang, G. Wen, Y. Hu, M. Luo, P. Yang, and Y. Xu. Mvanet: Multi-task guided multi-view attention network for chinese food recognition. *IEEE Transactions on Multimedia*, 23:3551–3561, 2021. ISSN 1941-0077. doi: 10.1109/TMM.2020.3028478.

[141] C. H. Lim, K. M. Goh, and L. L. Lim. Explainable artificial intelligence in oriental food recognition using convolutional neural network. In *2021 IEEE 11th International Conference on System Engineering and Technology (ICSET)*, pages 218–223, 2021. doi: 10.1109/ICSET53708.2021.9612442.

[142] Z. C. Lipton. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue*, 16(3):31–57, 2018. ISSN 1542-7730, 1542-7749. doi: 10.1145/3236386.3241340.

[143] F. T. Liu, K. M. Ting, and Z.-H. Zhou. Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422. IEEE, 2008. doi: 10.1109/ICDM.2008.17.

[144] Y. Liu and C. Aldrich. Anomaly detection and explanation in coal data using isolation forest, random forest, and shap. *International Journal of Coal Geology*, 250:103921, 2023. doi: 10.1016/j.coal.2023.103921.

[145] L. Longo, M. Brcic, F. Cabitza, J. Choi, R. Confalonieri, J. D. Ser, R. Guidotti, Y. Hayashi, F. Herrera, A. Holzinger, R. Jiang, H. Khosravi, F. Lecue, G. Malgieri, A. Páez, W. Samek, J. Schneider, T. Speith, and S. Stumpf. Explainable artificial intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106:102301, 2024. ISSN 1566-2535. doi: 10.1016/j.inffus.2024.102301.

[146] J. F. Lopes, V. G. T. da Costa, D. F. Barbin, L. J. P. Cruz-Tirado, V. Baeten, and S. Barbon Junior. Deep computer vision system for cocoa classification. *Multimedia Tools and Applications*, 81(28):41059–41077, 2022. ISSN 1573-7721. doi: 10.1007/s11042-022-13097-3.

[147] L. Lu, W. Liu, W. Yang, M. Zhao, and T. Jiang. Lightweight corn seed disease identification method based on improved shufflenetv2. *Agriculture*, 12 (11):1929, 2022. ISSN 2077-0472. doi: 10.3390/agriculture12111929.

[148] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 4768–4777, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

[149] M. Luo, W. Min, Z. Wang, J. Song, and S. Jiang. Ingredient prediction via context learning network with class-adaptive asymmetric loss. *IEEE Transactions on Image Processing*, 32:5509–5523, 2023. ISSN 1941-0042. doi: 10.1109/TIP.2023.3318958.

[150] P. Ma, C. P. Lau, N. Yu, A. Li, P. Liu, Q. Wang, and J. Sheng. Image-based nutrient estimation for chinese dishes using deep learning. *Food Research International*, 147:110437, 2021. ISSN 0963-9969. doi: 10.1016/j.foodres.2021.110437.

[151] P. Ma, C. P. Lau, N. Yu, A. Li, and J. Sheng. Application of deep learning for image-based chinese market food nutrients estimation. *Food Chemistry*, 373: 130994, 2022. ISSN 0308-8146. doi: 10.1016/j.foodchem.2021.130994.

[152] P. Ma, X. Jia, W. Xu, Y. He, K. Tarwa, M. O. Alharbi, C.-I. Wei, and Q. Wang. Enhancing salmon freshness monitoring with sol-gel cellulose nanocrystal colorimetric paper sensors and deep learning methods. *Food Bioscience*, 56: 103313, 2023. ISSN 2212-4292. doi: 10.1016/j.fbio.2023.103313.

[153] T. Mahmud, N. Datta, R. Chakma, U. Kanti Das, M. Tarek Aziz, M. Islam, A. Hasnat Muhammed Salimullah, M. Shahadat Hossain, and K. Andersson. An approach for crop prediction in agriculture: Integrating genetic algorithms and machine learning. *IEEE Access*, 12:173583–173598, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3478739.

[154] G. Makridis, E. Heyrman, D. Kotios, P. Mavrepis, B. Callens, R. V. De Vijver, J. Maselyne, M. Aluwe, and D. Kyriazis. Evaluating machine learning techniques to define the factors related to boar taint. *Livestock Science*, 264: 105045, 2022. ISSN 18711413. doi: 10.1016/j.livsci.2022.105045.

[155] G. Makridis, V. Koukos, G. Fatouros, M. M. Separdani, and D. Kyriazis. Enhancing explainability in mobility data science through a combination of methods. In *Intelligent Computing*, pages 45–60. Springer, 2024. ISBN 978-3-031-62269-4. doi: 10.1007/978-3-031-62269-4_4.

[156] I. Malashin, V. Tynchenko, A. Gantimurov, V. Nelyub, A. Borodulin, and Y. Tynchenko. Predicting sustainable crop yields: Deep learning and explainable AI tools. *Sustainability*, 16(21):9437, 2024. ISSN 2071-1050. doi: 10.3390/su16219437.

[157] L. J. Malcolmson and J. K. Winkler-Moser. Flavor and sensory aspects. *Bailey's Industrial Oil and Fat Products*, pages 1–17, 2020. doi: 10.1002/047167849X.bio032.pub2.

[158] A. Malekloo, E. Ozer, M. AlHamaydeh, and M. Girolami. Machine learning and structural health monitoring overview with emerging technology and high-dimensional data source highlights. *Structural Health Monitoring*, 21(4):1906–1955, 2022. doi: 10.1177/14759217211036880.

[159] L. Manning, S. Brewer, P. J. Craigon, J. Frey, A. Gutierrez, N. Jacobs, S. Kanza, S. Munday, J. Sacks, and S. Pearson. Artificial intelligence and ethics within the food sector: Developing a common language for technology adoption across the supply chain. *Trends in Food Science & Technology*, 125:33–42, 2022. ISSN 0924-2244. doi: 10.1016/j.tifs.2022.04.025.

[160] M. Marsot, J. Mei, X. Shan, L. Ye, P. Feng, X. Yan, C. Li, and Y. Zhao. An adaptive pig face recognition approach using convolutional neural networks. *Computers and Electronics in Agriculture*, 173:105386, 2020. ISSN 0168-1699. doi: 10.1016/j.compag.2020.105386.

[161] N. Martinel, G. L. Foresti, and C. Micheloni. Wide-slice residual networks for food recognition. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 567–576, 2018. doi: 10.1109/WACV.2018.00068.

[162] S. S. Martinelli and S. B. Cavalli. Healthy and sustainable diet: a narrative review of the challenges and perspectives. *Ciencia & Saude Coletiva*, 24(11):4251–4262, 2019. ISSN 1678-4561. doi: 10.1590/1413-812320182411.30572017.

[163] M. Mashayekhi and R. Gras. Rule extraction from random forest: the RF+HC methods. In D. Barbosa and E. Milios, editors, *Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 223–237. Springer International Publishing, 2015. ISBN 978-3-319-18356-5. doi: 10.1007/978-3-319-18356-5_20.

[164] A. Mateo-Sanchis, J. E. Adsuara, M. Piles, J. Munoz-Marí, A. Perez-Suay, and G. Camps-Valls. Interpretable long short-term memory networks for crop yield estimation. *IEEE Geoscience and Remote Sensing Letters*, 20:1–5, 2023. ISSN 1545-598X, 1558-0571. doi: 10.1109/LGRS.2023.3244064.

[165] R. M. Math and N. V. Dharwadkar. Early detection and identification of grape diseases using convolutional neural networks. *Journal of Plant Diseases and Protection*, 129(3):521–532, 2022. ISSN 1861-3837. doi: 10.1007/s41348-022-00589-5.

[166] R. Maurya, N. N. Pandey, V. P. Singh, and T. Gopalakrishnan. Plant disease classification using interpretable vision transformer network. *2023 International Conference on Recent Advances in Electrical, Electronics & Digital Healthcare Technologies (REEDCON)*, pages 688–692, 2023. doi: 10.1109/REEDCON57544.2023.10151342.

[167] R. Maurya, A. Srivastava, A. Srivastava, V. K. Pathak, and M. K. Dutta. Computer aided detection of mercury heavy metal intoxicated fish: an application of machine vision and artificial intelligence technique. *Multimedia Tools and Applications*, 82(13):20517–20536, 2023. ISSN 1573-7721. doi: 10.1007/s11042-023-14358-5.

[168] M. H. K. Mehedi, N. Nawer, S. Ahmed, M. S. I. Khan, K. M. Hasib, M. F. Mridha, M. G. R. Alam, and T. T. Nguyen. PLD-det: plant leaf disease detection in real time using an end-to-end neural network approach based on improved YOLOv7. *Neural Computing and Applications*, 36(34):21885–21898, 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-10409-6.

[169] L. Meng, F. Feng, X. He, X. Gao, and T.-S. Chua. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. *Proceedings of the 28th ACM International Conference on Multimedia*, pages 3460–3468, 2020. doi: 10.1145/3394171.3413598.

[170] Z. Mi, X. Zhang, J. Su, D. Han, and B. Su. Wheat stripe rust grading by deep learning with attention mechanism and images from mobile devices. *Frontiers in Plant Science*, 11, 2020. ISSN 1664-462X. doi: 10.3389/fpls.2020.558126.

[171] I. D. Mienye and Y. Sun. A survey of ensemble learning: Concepts, algorithms, applications, and prospects. *IEEE Access*, 10:99129–99149, 2022. ISSN 2169-3536. doi: 10.1109/ACCESS.2022.3207287.

[172] W. Min, Z. Wang, J. Yang, C. Liu, and S. Jiang. Vision-based fruit recognition via multi-scale attention CNN. *Computers and Electronics in Agriculture*, 210: 107911, 2023. ISSN 0168-1699. doi: 10.1016/j.compag.2023.107911.

[173] R. Mishra, A. Kavita, Rajpal, V. Bhatia, S. Rajpal, M. Agarwal, and N. Kumar. I-ldd: an interpretable leaf disease detector. *Soft Computing*, 2023. ISSN 1433-7479. doi: 10.1007/s00500-023-08512-2.

[174] Y. Miura, Y. Sawamura, Y. Shinomiya, and S. Yoshida. Vegetable mass estimation based on monocular camera using convolutional neural network. *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2106–2112, 2020. doi: 10.1109/SMC42975.2020.9282930.

[175] R. N. V. J. Mohan, P. S. Rayanoothala, and R. P. Sree. Next-gen agriculture: integrating AI and XAI for precision crop yield predictions. *Frontiers in Plant Science*, 15, 2025. ISSN 1664-462X. doi: 10.3389/fpls.2024.1451607.

[176] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of Machine Learning*. Mit Pr, Cambridge, MA, 2012. ISBN 978-0-262-01825-8.

[177] E. Mones, L. Vicsek, and T. Vicsek. Hierarchy measure for complex networks. *PloS one*, 7(3):e33799, 2012. ISSN 1932-6203. doi: 10.1371/journal.pone.0033799.

[178] I. A. Moraes, L. Arrighi, S. B. Junior, R. L. Cunha, and D. F. Barbin. Explainable artificial intelligence (XAI) applied to deep computer vision for the assessment and classification of oleogels with varying oleogelator types and concentrations. *Microchemical Journal*, page 116821, 2026. ISSN 0026-265X. doi: 10.1016/j.microc.2026.116821.

[179] M. S. Morshed, S. Ahmed, T. Ahmed, M. U. Islam, and A. Ashikur Rahman. Fruit quality assessment with densely connected convolutional neural network. In *2022 12th International Conference on Electrical and Computer Engineering (ICECE)*, pages 1–4, 2022. doi: 10.1109/ICECE57408.2022.10088873.

[180] R. R. Mukhametzyanov, S. V. Brusenko, A. M. Khezhev, E. M. Kelemetov, and S. S. Kirillova. Changing the global production and trade of citrus fruits. In *Sustainable Development of the Agrarian Economy Based on Digital Technologies and Smart Innovations*, pages 19–24. Springer, 2024. doi: 10.1007/978-3-031-51272-8_4.

[181] K. P. Murphy. *Probabilistic Machine Learning: An Introduction*. MIT Press, 2022.

[182] A. Murtovi, A. Bainczyk, G. Nolte, M. Schlüter, and B. Steffen. Forest GUMP: a tool for verification and explanation. *International Journal on Software Tools for Technology Transfer*, 25(3):287–299, 2023. doi: 10.1007/s10009-023-00702-5.

[183] M. Mustak Un Nobi, M. Rifat, M. F. Mridha, S. Alfarhood, M. Safran, and D. Che. Gld-det: Guava leaf disease detection in real-time using lightweight deep learning approach based on mobilenet. *Agronomy*, 13(9):2240, 2023. ISSN 2073-4395. doi: 10.3390/agronomy13092240.

[184] P. Naga Srinivasu, M. F. Ijaz, and M. Woźniak. XAI-driven model for crop recommender system for use in precision agriculture. *Computational Intelligence*, 40(1):e12629, 2024. ISSN 1467-8640. doi: 10.1111/coin.12629.

[185] B. B. Nair and V. P. Mohandas. Artificial intelligence applications in financial forecasting—a survey and some empirical results. *Int. Dec. Tech.*, 9(2):99–140, 2015. ISSN 1872-4981.

[186] H. Nakahara, A. Jinguji, S. Sato, and T. Sasao. A random forest using a multi-valued decision diagram on an FPGA. In *2017 IEEE 47th International Symposium on Multiple-Valued Logic (ISMVL)*, pages 266–271, 2017. doi: 10.1109/ISMVL.2017.40.

[187] M.-L. Ndao, G. Youness, N. Niang, and G. Saporta. Enhancing explainability in predictive maintenance: Investigating the impact of data preprocessing techniques on xai effectiveness. In *The 37th International Conference of the*

Florida Artificial Intelligence Research Society, Florida, United States, 2024. doi: 10.32473/flairs.37.1.135526.

[188] S. Needham and D. L. Dowe. Message length as an effective ockham's razor in decision tree induction. In *International Workshop on Artificial Intelligence and Statistics*, pages 216–223. PMLR, 2001.

[189] H. Nematzadeh, J. García-Nieto, S. Hurtado, J. F. Aldana-Montes, and I. Navas-Delgado. Model-agnostic local explanation: Multi-objective genetic algorithm explainer. *Engineering Applications of Artificial Intelligence*, 139: 109628, 2025. ISSN 0952-1976. doi: 10.1016/j.engappai.2024.109628.

[190] M. P. Neto and F. V. Paulovich. Explainable matrix - visualization for global and local interpretability of random forest classification ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1427–1437, 2020. ISSN 1077-2626. doi: 10.1109/TVCG.2020.3030354.

[191] J. Ni, Y. Zhao, Z. Zhou, L. Zhao, and Z. Han. Condiment recognition using convolutional neural networks with attention mechanism. *Journal of Food Composition and Analysis*, 115:104964, 2023. ISSN 0889-1575. doi: 10.1016/j.jfca.2022.104964.

[192] N. Nigar, H. Muhammad Faisal, M. Umer, O. Oki, and J. Manappattukunnel Lukose. Improving plant disease classification with deep-learning-based prediction model using explainable artificial intelligence. *IEEE Access*, 12: 100005–100014, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3428553.

[193] S. F. Nimmy, M. S. Kamal, O. K. Hussain, and R. Chakrab. Interpretability in mapping weeds and crops from drone images. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8, 2024. doi: 10.1109/IJCNN60899.2024.10650761.

[194] S. M. N. Nobel, M. Afroj, M. M. Kabir, and M. F. Mridha. Development of a cutting-edge ensemble pipeline for rapid and accurate diagnosis of plant leaf diseases. *Artificial Intelligence in Agriculture*, 14:56–72, 2024. ISSN 2589-7217. doi: 10.1016/j.aiia.2024.10.005.

[195] R. Nomura and K. Oki. Development of health monitoring method for pecan nut trees using side video data and computer vision. *Optical Review*, 28(6): 730–737, 2021. ISSN 1349-9432. doi: 10.1007/s10043-021-00694-0.

[196] P. Novielli, D. Romano, S. Pavan, P. Losciale, A. M. Stellacci, D. Diacono, R. Bellotti, and S. Tangaro. Explainable artificial intelligence for genotype-to-phenotype prediction in plant breeding: a case study with a dataset from an almond germplasm collection. *Frontiers in Plant Science*, 15, 2024. ISSN 1664-462X. doi: 10.3389/fpls.2024.1434229.

[197] V. Nussiri and P. Vateekul. Food image categorization using attentional bilinear model. In *2019 11th International Conference on Information Technology and Electrical Engineering (ICITEE)*, pages 1–6, 2019. doi: 10.1109/ICITEED.2019.8929982.

[198] A. Oad, S. S. Abbas, A. Zafar, B. A. Akram, F. Dong, M. S. H. Talpur, and M. Uddin. Plant leaf disease detection using ensemble learning and explainable AI. *IEEE Access*, 12:156038–156049, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3484574.

[199] A. G. Olenskyj, I. R. Donis-González, J. M. Earles, and G. M. Bornhorst. End-to-end prediction of uniaxial compression profiles of apples during in vitro digestion using time-series micro-computed tomography and deep learning. *Journal of Food Engineering*, 325:111014, 2022. ISSN 0260-8774. doi: 10.1016/j.jfoodeng.2022.111014.

[200] R. C. d. Oliveira and R. D. d. S. e. Silva. Artificial intelligence in agriculture: Benefits, challenges, and trends. *Applied Sciences*, 13(13):7405, 2023. ISSN 2076-3417. doi: 10.3390/app13137405.

[201] J. Oliver. *Decision Graphs - An Extension of Decision Trees*. Monash University, Department of Computer Science, 1992.

[202] D. O'Rourke. Economic importance of the world apple industry. In S. S. Korban, editor, *The Apple Genome*, pages 1–18. Springer International Publishing, 2021. doi: 10.1007/978-3-030-74682-7_1.

[203] S. Othman, N. R. Mavani, M. A. Hussain, N. A. Rahman, and J. Mohd Ali. Artificial intelligence-based techniques for adulteration and defect detections in food and agricultural industry: A review. *Journal of Agriculture and Food Research*, 12:100590, 2023. ISSN 2666-1543. doi: 10.1016/j.jafr.2023.100590.

[204] P. Patil, S. K. Pamali, S. B. Devagiri, A. S. Sushma, and J. Mirje. Plant leaf disease detection using XAI. In *2024 3rd International Conference on Artificial Intelligence For Internet of Things (AIIoT)*, pages 1–6, 2024. doi: 10.1109/AIIoT58432.2024.10574617.

[205] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[206] L. F. S. Pereira, S. Barbon Jr, N. A. Valous, and D. F. Barbin. Predicting the ripening of papaya fruit with digital imaging and random forests. *Computers and Electronics in Agriculture*, 145:76–82, 2018. doi: 10.1016/j.compag.2017.12.029.

[207] C. Peri. The universe of food quality. *Food Quality and Preference*, 17(1):3–8, 2006. ISSN 0950-3293. doi: 10.1016/j.foodqual.2005.03.002.

[208] R. P. Porfirio, P. A. Santos, and R. N. Madeira. Enhancing digital agriculture with XAI: Case studies on tabular data and future directions. *Companion Proceedings of the 26th International Conference on Multimodal Interaction*, pages 211–217, 2024. doi: 10.1145/3686215.3689201.

[209] R. P. Porfírio, R. N. Madeira, and P. A. Santos. AgriUXE: Integrating explainable AI and multimodal data for smart agriculture. In *2024 International Symposium on Sensing and Instrumentation in 5G and IoT Era (ISSI)*, volume 1, pages 1–6, 2024. doi: 10.1109/ISSI63632.2024.10720487.

[210] B. Prashanthi, A. V. P. Krishna, and C. M. Rao. LEViT- leaf disease identification and classification using an enhanced vision transformers(ViT) model. *Multimedia Tools and Applications*, 2024. ISSN 1573-7721. doi: 10.1007/s11042-024-19866-6.

[211] V. K. Pratap and N. S. Kumar. High-precision multiclass classification of chili leaf disease through customized efficientnetb4 from chili leaf images. *Smart Agricultural Technology*, 5:100295, 2023. ISSN 2772-3755. doi: 10.1016/j.atech.2023.100295.

[212] A. Preece. Asking 'Why' in AI: Explainability of intelligent systems – perspectives and challenges. *Intelligent Systems in Accounting, Finance and Management*, 25(2):63–72, 2018. ISSN 2160-0074. doi: 10.1002/isaf.1422.

[213] K. Przybył. Explainable AI: Machine learning interpretation in blackcurrant powders. *Sensors*, 24(10):3198, 2024. ISSN 1424-8220. doi: 10.3390/s24103198.

[214] L.-D. Quach, K. N. Quoc, A. N. Quynh, H. T. Ngoc, and N. Thai-Nghe. Tomato health monitoring system: Tomato classification, detection, and counting system based on YOLOv8 model with explainable MobileNet models using grad-CAM++. *IEEE Access*, 12:9719–9737, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3351805.

[215] L. Rachwał, B. Krawczyk, and M. Woźniak. Isolation forest with exclusion of attributes based on shapley index. *IEEE Transactions on Neural Networks and Learning Systems*, 34(8):4011–4022, 2023. doi: 10.1109/ACCESS.2024.3432174.

[216] F. Radicchi, C. Castellano, F. Cecconi, V. Loreto, and D. Parisi. Defining and identifying communities in networks. In *Proceedings of the National Academy of Sciences*, volume 101, pages 2658–2663, 2004. doi: 10.1073/pnas.0400054101.

[217] U. N. Raghavan, R. Albert, and S. Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76 (3):036106, 2007. doi: 10.1103/PhysRevE.76.036106.

[218] M. M. Rahman, Z. Yan, M. T. Aziz, M. A. B. Siddick, T. Truong, M. M. Sharif, N. Datta, T. Mahmud, R. D. C. Pecho, and S. M. Farid. Explainable deep transfer learning framework for rice leaf disease diagnosis and classification. *International Journal of Advanced Computer Science and Applications (ijacsa)*, 15(12), 2024. ISSN 2156-5570. doi: 10.14569/IJACSA.2024.0151287.

[219] D. Ramírez-Mejía, C. Levers, and J.-F. Mas. Spatial patterns and determinants of avocado frontier dynamics in mexico. *Regional Environmental Change*, 22 (1):28, 2022. ISSN 1436-378X. doi: 10.1007/s10113-022-01883-6.

[220] M. S. Rashid, M. S. Morshed, M. U. Islam, S. Rashid, A. Mahmud, and A. Islam. Mycological examination of microscopic fungi images with deep learning and gradient weighted class activation mapping visualization. In *2024 Advances in Science and Engineering Technology International Conferences (ASET)*, pages 01–08, 2024. doi: 10.1109/ASET60340.2024.10708690.

[221] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 779–788, Las Vegas, NV, USA, 2016. IEEE. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.91.

[222] E. Rho, M. Kim, S. H. Cho, B. Choi, H. Park, H. Jang, Y. S. Jung, and S. Jo. Separation-free bacterial identification in arbitrary media via deep neural network-based SERS analysis. *Biosensors and Bioelectronics*, 202:113991, 2022. ISSN 0956-5663. doi: 10.1016/j.bios.2022.113991.

[223] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 1135–1144. Association for Computing Machinery, 2016. ISBN 978-1-4503-4232-2. doi: 10.1145/2939672.2939778.

[224] L. Rizzo and L. Longo. A qualitative investigation of the explainability of defeasible argumentation and non-monotonic fuzzy reasoning. In *Proceedings for the 26th AIAI Irish Conference on Artificial Intelligence and Cognitive Science Trinity College Dublin, Dublin, Ireland, December 6-7th, 2018.*, pages 138–149, 2018. doi: https://doi.org/10.21427/tby8-8z04.

[225] H. Rogers, B. De La Iglesia, T. Zebin, G. Cielniak, and B. Magri. Advancing precision agriculture: domain-specific augmentations and robustness testing for convolutional neural networks in precision spraying evaluation. *Neural Computing and Applications*, 36(32):20211–20229, 2024. ISSN 1433-3058. doi: 10.1007/s00521-024-10142-0.

[226] A. Roos. The european union's general data protection regulation (GDPR) and its implications for south african data privacy law: An evaluation of selected 'content principles'. *The Comparative and International Law Journal of Southern Africa*, 53(3):1–37, 2020. ISSN 0010-4051.

[227] M. Rostami, U. Muhammad, S. Forouzandeh, K. Berahmand, V. Farrahi, and M. Oussalah. An effective explainable food recommendation using deep image clustering and community detection. *Intelligent Systems with Applications*, 16: 200157, 2022. ISSN 2667-3053. doi: 10.1016/j.iswa.2022.200157.

[228] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel distributed processing: explorations in the microstructure of cognition, vol. 1: foundations*, pages 318–362. MIT Press, Cambridge, MA, USA, 1986. ISBN 978-0-262-68053-0.

[229] M. Ryo. Explainable artificial intelligence and interpretable machine learning for agricultural data analysis. *Artificial Intelligence in Agriculture*, 6:257–265, 2022. ISSN 2589-7217. doi: 10.1016/j.aiia.2022.11.003.

[230] F. Sabrina, S. Sohail, F. Farid, S. Jahan, F. Ahamed, and S. Gordon. An interpretable artificial intelligence based smart agriculture system. *Computers, Materials & Continua*, pages 3777–3797, 2022. ISSN 1546-2218. doi: 10.32604/cmc.2022.026363.

[231] A. Salam, M. Naznine, N. Jahan, E. Nahid, M. Nahiduzzaman, and M. E. H. Chowdhury. Mulberry leaf disease detection using CNN-based smart android application. *IEEE Access*, 12:83575–83588, 2024. ISSN 2169-3536. doi: 10.1109/ACCESS.2024.3407153.

[232] G. I. Sayed and A. E. Hassanien. Explainable ai and slime mould algorithm for classification of pistachio species. *Artificial Intelligence: A Real Opportunity in the Food Industry*, pages 29–43, 2023. doi: 10.1007/978-3-031-13702-0_3.

[233] C. J. Seal and K. Brandt. 3 - nutritional quality of foods. *Handbook of Organic Food Safety and Quality*, pages 25–40, 2007. doi: 10.1533/9781845693411.1.25.

[234] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 618–626, 2017. doi: 10.1109/ICCV.2017.74. ISSN: 2380-7504.

[235] K. Sermmany, P. Wanjantuk, and W. Leelapatra. Utilizing explainable artificial intelligence (XAI) to identify determinants of coffee quality. In *2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 696–703, 2024. doi: 10.1109/JCSSE61278.2024.10613641. ISSN: 2642-6579.

[236] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti. GLocalX - from local to global explanations of black box AI models. *Artificial Intelligence*, 294:103457, 2021. ISSN 0004-3702. doi: 10.1016/j.artint.2021.103457.

[237] G. B. Seymour, M. Poole, J. J. Giovannoni, and G. A. Tucker. *The Molecular Biology and Biochemistry of Fruit Ripening*. Wiley, 2013. doi: 10.1002/9781118593714.

[238] F. Shahoveisi, H. Taheri Gorji, S. Shahabi, S. Hosseinirad, S. Markell, and F. Vasefi. Application of image processing and transfer learning for the detection of rust disease. *Scientific Reports*, 13(1):5133, 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-31942-9.

[239] M. K. Shaik, M. L. Prasad, Y. S. Reddy, S. Asif, D. Kalpana, and P. C. S. Reddy. Smart agriculture: Explainable deep learning approach with gradient-weighted class activation mapping. In *2024 International Conference on Computer, Electronics, Electrical Engineering & their Applications (IC2E3)*, pages 1–6, 2024. doi: 10.1109/IC2E362166.2024.10826675.

[240] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning From Theory to Algorithms*. Cambridge University Press, New York, NY, USA, 2014. ISBN 978-1-107-05713-5.

[241] M. Y. Shams, S. A. Gamel, and F. M. Talaat. Enhancing crop recommendation systems with explainable artificial intelligence: a study on agricultural decision-making. *Neural Computing and Applications*, 36(11):5695–5714, 2024. ISSN 1433-3058. doi: 10.1007/s00521-023-09391-2.

[242] W. Shao, S. Hou, W. Jia, and Y. Zheng. Rapid non-destructive analysis of food nutrient content using swin-nutrition. *Foods*, 11(21):3429, 2022. ISSN 2304-8158. doi: 10.3390/foods11213429.

[243] W. Shao, W. Min, S. Hou, M. Luo, T. Li, Y. Zheng, and S. Jiang. Vision-based food nutrition estimation via RGB-d fusion network. *Food Chemistry*, 424:136309, 2023. ISSN 0308-8146. doi: 10.1016/j.foodchem.2023.136309.

[244] R. Shi, T. Li, L. Zhang, and Y. Yamaguchi. Visualization comparison of vision transformers and convolutional neural networks. *IEEE Transactions on Multimedia*, 26:2327–2339, 2024. doi: 10.1109/TMM.2023.3294805.

[245] S. Shin, Y. Lee, S. Kim, S. Choi, J. G. Kim, and K. Lee. Rapid and non-destructive spectroscopic method for classifying beef freshness using a deep spectral network fused with myoglobin information. *Food Chemistry*, 352: 129329, 2021. ISSN 0308-8146. doi: 10.1016/j.foodchem.2021.129329.

[246] M. Shoaib, T. Hussain, B. Shah, I. Ullah, S. M. Shah, F. Ali, and S. H. Park. Deep learning-based segmentation and classification of leaf images for detection of tomato plant disease. *Frontiers in Plant Science*, 13, 2022. ISSN 1664-462X. doi: 10.3389/fpls.2022.1031748.

[247] T. E. Shrestha, A. R. Aurnob, S. A. Tanim, M. Islam, and K. Nur. Revolutionizing cucumber agriculture: AI for precision classification of leaf diseases. In *2024 6th International Conference on Electrical Engineering and Information & Communication Technology (ICEEICT)*, pages 776–781, 2024. doi: 10.1109/ICEEICT62016.2024.10534530. ISSN: 2769-5700.

[248] A. Shrikumar, P. Greenside, and A. Kundaje. Learning important features through propagating activation differences. *Proceedings of the 34th International Conference on Machine Learning*, pages 3145–3153, 2017.

[249] D. Sihi, B. Dari, A. P. Kuruvila, G. Jha, and K. Basu. Explainable machine learning approach quantified the long-term (1981–2015) impact of climate and soil properties on yields of major agricultural crops across conus. *Frontiers in Sustainable Food Systems*, 6, 2022. ISSN 2571-581X. doi: 10.3389/fsufs.2022.847892.

[250] O. Silva, A. Silva, I. Moreira, J. Nacif, and R. Ferreira. RDSF: Everything at same place all at once - a random decision single forest. In *Anais do XIII Simpósio Brasileiro de Engenharia de Sistemas Computacionais*, 2023.

[251] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[252] N. Singh and M. Adhikari. Real-time paddy field irrigation using feature extraction and federated learning strategy. *IEEE Sensors Journal*, 24(21): 36159–36166, 2024. ISSN 1558-1748. doi: 10.1109/JSEN.2024.3462496.

[253] J. Sipple and A. Youssef. A general-purpose method for applying explainable AI for anomaly detection. In *International Symposium on Methodologies for Intelligent Systems*, pages 162–174. Springer, 2022. ISBN 978-3-031-16564-1. doi: 10.1007/978-3-031-16564-1_16.

[254] T. Speith. A review of taxonomies of explainable artificial intelligence (XAI) methods. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 2239–2250. Association for Computing Machinery, 2022. ISBN 978-1-4503-9352-2. doi: 10.1145/3531146.3534639.

[255] S. Strasser and M. Klettke. Transparent data preprocessing for machine learning. In *Proceedings of the 2024 Workshop on Human-In-the-Loop Data Analytics*, pages 1–6, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400706936. doi: 10.1145/3665939.3665960.

[256] C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8:25, 2007. ISSN 1471-2105. doi: 10.1186/1471-2105-8-25.

[257] G. A. Tahir and C. K. Loo. Explainable deep learning ensemble for food image analysis on edge devices. *Computers in Biology and Medicine*, 139: 104972, 2021. ISSN 0010-4825. doi: 10.1016/j.compbiomed.2021.104972.

[258] G. A. Tahir and C. K. Loo. Progressive kernel extreme learning machine for food image analysis via optimal features from quality resilient CNN. *Applied Sciences*, 11(20):9562, 2021. ISSN 2076-3417. doi: 10.3390/app11209562.

[259] P. J. Tan and D. L. Dowe. MML inference of decision graphs with multi-way joins and dynamic attributes. In T. T. D. Gedeon and L. C. C. Fung, editors, *AI 2003: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, pages 269–281. Springer, 2003. ISBN 978-3-540-24581-0. doi: 10.1007/978-3-540-24581-0_23.

[260] S. A. Tanim, T. E. Shrestha, K. Tanvir, M. S. Kabir, M. F. Mridha, and M. K. Haq. Single-level fusion for enhancing meat quality classification with explainable AI. In *2024 IEEE International Conference on Computing, Applications and Systems (COMPAS)*, pages 1–6, 2024. doi: 10.1109/COMPAS60761.2024.10796775.

[261] M. Tariq, U. Ali, S. Abbas, S. Hassan, R. A. Naqvi, M. A. Khan, and D. Jeong. Corn leaf disease: insightful diagnosis using VGG16 empowered by explainable

AI. *Frontiers in Plant Science*, 15, 2024. ISSN 1664-462X. doi: 10.3389/fpls.2024.1402835.

[262] C. Taylor, J. Guy, and J. Bacardit. Estimating individual-level pig growth trajectories from group-level weight time series using machine learning. *Computers and Electronics in Agriculture*, 208:107790, 2023. ISSN 0168-1699. doi: 10.1016/j.compag.2023.107790.

[263] A. Tempelaere, H. Minh Phan, T. van de Looverbosch, P. Verboven, and B. Nicolai. Non-destructive internal disorder segmentation in pear fruit by x-ray radiography and ai. *Computers and Electronics in Agriculture*, 212:108142, 2023. ISSN 0168-1699. doi: 10.1016/j.compag.2023.108142.

[264] A. Tempelaere, L. Van Doorselaer, J. He, P. Verboven, and B. M. Nicolai. Braenet: Internal disorder detection in 'braeburn' apple using x-ray imaging data. *Food Control*, 155:110092, 2024. ISSN 0956-7135. doi: 10.1016/j.foodcont.2023.110092.

[265] K. Terada and K. Fujinami. Improving disease forecast on different farms using sensing agricultural robot with XAI. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 398–402, 2024. doi: 10.1109/GCCE62371.2024.10760700.

[266] D. M. Thomas, S. Kleinberg, A. W. Brown, M. Crow, N. D. Bastian, N. Reisweber, R. Lasater, T. Kendall, P. Shafto, R. Blaine, S. Smith, D. Ruiz, C. Morrell, and N. Clark. Machine learning modeling practices to support the principles of AI and ethics in nutrition research. *Nutrition & Diabetes*, 12(1):1–10, 2022. ISSN 2044-4052. doi: 10.1038/s41387-022-00226-y.

[267] J. Torres-Tello and S.-B. Ko. Interpretability of artificial intelligence models that use data fusion to predict yield in aeroponics. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3331–3342, 2023. ISSN 1868-5145. doi: 10.1007/s12652-021-03470-9.

[268] N. L. Tsakiridis, T. Diamantopoulos, A. L. Symeonidis, J. B. Theocharis, A. Iossifides, P. Chatzimisios, G. Pratos, and D. Kouvas. Versatile internet of things for agriculture: An eXplainable AI approach. *Artificial Intelligence Applications and Innovations*, pages 180–191, 2020. doi: 10.1007/978-3-030-49186-4_16.

[269] J. L. Valenzuela. Advances in postharvest preservation and quality of fruits and vegetables. *Foods*, 12(9):1830, 2023. ISSN 2304-8158. doi: 10.3390/foods12091830.

[270] A. Van Assche and H. Blockeel. Seeing the forest through the trees: Learning a comprehensible model from an ensemble. In J. N. Kok, J. Koronacki, R. L. d. Mantaras, S. Matwin, D. Mladenič, and A. Skowron, editors, *Machine Learning: ECML 2007*, pages 418–429. Springer, 2007. ISBN 978-3-540-74958-5. doi: 10.1007/978-3-540-74958-5_39.

[271] T. van de Looverbosch, J. He, A. Tempelaere, K. Kelchtermans, P. Verboven, T. Tuytelaars, J. Sijbers, and B. Nicolai. Inline nondestructive internal disorder detection in pear fruit using explainable deep anomaly detection on x-ray images. *Computers and Electronics in Agriculture*, 197:106962, 2022. ISSN 0168-1699. doi: 10.1016/j.compag.2022.106962.

[272] N. J. Vardhan, D. Chandana, R. Dheepak Raaj, S. Shanmukhi, and A. Radhakrishnan. A comparative study of hyperparameter tuning in deep learning models using bayesian optimization and XAI. In *2024 15th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–6, 2024. doi: 10.1109/ICCCNT61001.2024.10725868.

[273] C. M. Viana, M. Santos, D. Freire, P. Abrantes, and J. Rocha. Evaluation of the factors explaining the use of agricultural land: A machine learning and model-agnostic approach. *Ecological Indicators*, 131:108200, 2021. ISSN 1470-160X. doi: 10.1016/j.ecolind.2021.108200.

[274] G. Vilone and L. Longo. Classification of explainable artificial intelligence methods through their output formats. *Machine Learning and Knowledge Extraction*, 3(3):615–661, 2021. ISSN 2504-4990. doi: 10.3390/make3030032.

[275] G. Vilone and L. Longo. Development of a human-centred psychometric test for the evaluation of explanations produced by XAI methods. In *Explainable Artificial Intelligence*, pages 205–232. Springer Nature Switzerland, 2023. ISBN 978-3-031-44070-0. doi: 10.1007/978-3-031-44070-0_11.

[276] G. Vinci, R. Preti, A. Tieri, and S. Vieri. Authenticity and quality of animal origin food investigated by stable-isotope ratio analysis. *Journal of the Science of Food and Agriculture*, 93(3):439–448, 2013. ISSN 1097-0010. doi: 10.1002/jsfa.5970.

[277] S. A. Wadood, G. Boli, Z. Xiaowen, I. Hussain, and W. Yimin. Recent development in the application of analytical techniques for the traceability and authenticity of food of plant origin. *Microchemical Journal*, 152:104295, 2020. ISSN 0026-265X. doi: 10.1016/j.microc.2019.104295.

[278] C. S. Wallace. *Statistical and Inductive Inference by Minimum Message Length*. Information Science and Statistics. Springer Science & Business Media, 2005. ISBN 978-0-387-23795-4. doi: 10.1007/0-387-27656-4.

[279] C. Wang, D. Caragea, N. Kodadinne Narayana, N. T. Hein, R. Bheemanahalli, I. M. Somayanda, and S. V. K. Jagadish. Deep learning based high-throughput phenotyping of chalkiness in rice exposed to high night temperature. *Plant Methods*, 18(1):9, 2022. ISSN 1746-4811. doi: 10.1186/s13007-022-00839-5.

[280] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: Score-weighted visual explanations for convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 24–25, 2020.

[281] Y. Wang, F. Zhu, C. J. Boushey, and E. J. Delp. Weakly supervised food image segmentation using class activation maps. *2017 IEEE International Conference on Image Processing (ICIP)*, pages 1277–1281, 2017. doi: 10.1109/ICIP.2017.8296487.

[282] Y. Wang, J. Chandrasekaran, F. Haberkorn, Y. Dong, M. Gopinath, and F. A. Batarseh. Deepfarm: Ai-driven management of farm production using explainable causality. In *2022 IEEE 29th Annual Software Technology Conference (STC)*, pages 27–36, 2022. doi: 10.1109/STC55697.2022.00013.

[283] K. Wei, B. Chen, J. Zhang, S. Fan, K. Wu, G. Liu, and D. Chen. Explainable deep learning study for leaf disease classification. *Agronomy*, 12(5):1035, 2022. ISSN 2073-4395. doi: 10.3390/agronomy12051035.

[284] S. Weng, K. Han, Z. Chu, G. Zhu, C. Liu, Z. Zhu, Z. Zhang, L. Zheng, and L. Huang. Reflectance images of effective wavelengths from hyperspectral imaging for identification of fusarium head blight-infected wheat kernels combined with a residual attention convolution neural network. *Computers and Electronics in Agriculture*, 190:106483, 2021. ISSN 0168-1699. doi: 10.1016/j.compag.2021.106483.

[285] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. The what-if tool: Interactive probing of machine learning models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):56–65, 2020. ISSN 1941-0506. doi: 10.1109/TVCG.2019.2934619.

[286] A. Wolanin, G. Mateo-García, G. Camps-Valls, L. Gómez-Chova, M. Meroni, G. Duveiller, Y. Liangzhi, and L. Guanter. Estimating and understanding crop yields with explainable deep learning in the indian wheat belt. *Environmental Research Letters*, 15(2):024019, 2020. ISSN 1748-9326. doi: 10.1088/1748-9326/ab68ac.

[287] W. Wu, Y. Su, X. Chen, S. Zhao, I. King, M. R. Lyu, and Y.-W. Tai. Towards global explanations of convolutional neural networks with concept attribution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8649–8658, 2020. doi: 10.1109/CVPR42600.2020.00868.

[288] X. Xie, Y. Ge, H. Walia, J. Yang, and H. Yu. Leaf-counting in monocot plants using deep regression models. *Sensors*, 23(4):1890, 2023. ISSN 1424-8220. doi: 10.3390/s23041890.

[289] L. Xu, S. Ning, X. Xu, S. Wang, L. Chen, R. Long, S. Zhang, Y. Zhou, M. Zhang, and B. R. Thapa. Comparative analysis of machine learning models and explainable AI for agriculture drought prediction: A case study of the tapieh mountains. *Agricultural Water Management*, 306:109176, 2024. ISSN 0378-3774. doi: 10.1016/j.agwat.2024.109176.

[290] P. Xu, Q. Tan, Y. Zhang, X. Zha, S. Yang, and R. Yang. Research on maize seed classification and recognition based on machine vision and deep

learning. *Agriculture*, 12(2):232, 2022. ISSN 2077-0472. doi: 10.3390/agriculture12020232.

[291] T. Yamaguchi, T. Takamura, T. S. T. Tanaka, T. Ookawa, and K. Katsura. A study on optimal input images for rice yield prediction models using CNN with UAV imagery and its reasoning using explainable AI. *European Journal of Agronomy*, 164:127512, 2025. ISSN 1161-0301. doi: 10.1016/j.eja.2025.127512.

[292] Y. Yang, G. Jiao, J. Liu, W. Zhao, and J. Zheng. A lightweight rice disease identification network based on attention mechanism and dynamic convolution. *Ecological Informatics*, 78:102320, 2023. ISSN 1574-9541. doi: 10.1016/j.ecoinf.2023.102320.

[293] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. GNNExplainer: Generating explanations for graph neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[294] J. Yu, W. Zheng, L. Xu, F. Meng, J. Li, and L. Zhangzhong. Tpe-catboost: An adaptive model for soil moisture spatial estimation in the main maize-producing areas of china with multiple environment covariates. *Journal of Hydrology*, 613:128465, 2022. ISSN 0022-1694. doi: 10.1016/j.jhydrol.2022.128465.

[295] Z. Yu, H. Fang, Q. Zhangjin, C. Mi, X. Feng, and Y. He. Hyperspectral imaging technology combined with deep learning for hybrid okra seed identification. *Biosystems Engineering*, 212:46–61, 2021. ISSN 1537-5110. doi: 10.1016/j.biosystemseng.2021.09.010.

[296] Z. Yuan, K. Liu, S. Li, and P. Yang. Automatic generation of visual concept-based explanations for pest recognition. *2023 IEEE 21st International Conference on Industrial Informatics (INDIN)*, pages 1–6, 2023. doi: 10.1109/INDIN51400.2023.10217975.

[297] C. V. G. Zelaya. Towards explaining the effects of data preprocessing on machine learning. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 2086–2090. IEEE, 2019. doi: 10.1109/ICDE.2019.00245.

[298] D. Zhan, Y. Mu, W. Duan, M. Ye, Y. Song, Z. Song, K. Yao, D. Sun, and Z. Ding. Spatial prediction and mapping of soil water content by TPE-GBDT model in chinese coastal delta farmland with sentinel-2 remote sensing data. *Agriculture*, 13(5):1088, 2023. ISSN 2077-0472. doi: 10.3390/agriculture13051088.

[299] G. Zhang and W. Abdulla. Explainable ai-driven wavelength selection for hyperspectral imaging of honey products. *Food Chemistry Advances*, 3:100491, 2023. ISSN 2772-753X. doi: 10.1016/j.focha.2023.100491.

[300] J. Zhang, D. Lee, K. Jungles, D. Shaltis, K. Najarian, R. Ravikumar, G. Sanders, and J. Gryak. Prediction of oral food challenge outcomes via ensemble learning. *Informatics in Medicine Unlocked*, 36:101142, 2023. ISSN 2352-9148. doi: 10.1016/j.imu.2022.101142.

[301] X. Zhang, H. Gao, and L. Wan. Classification of fine-grained crop disease by dilated convolution and improved channel attention module. *Agriculture*, 12 (10):1727, 2022. ISSN 2077-0472. doi: 10.3390/agriculture12101727.

[302] Y. Zhang, C. Wei, Y. Zhong, H. Wang, H. Luo, and Z. Weng. Deep learning detection of shrimp freshness via smartphone pictures. *Journal of Food Measurement and Characterization*, 16(5):3868–3876, 2022. ISSN 2193-4134. doi: 10.1007/s11694-022-01473-4.

[303] H. Zhao, K.-H. Yap, A. C. Kot, and L. Duan. Jdnet: A joint-learning distilled network for mobile visual food recognition. *IEEE Journal of Selected Topics in Signal Processing*, 14(4):665–675, 2020. ISSN 1941-0484. doi: 10.1109/JSTSP.2020.2969328.

[304] X. Zhao, Y. Wu, D. L. Lee, and W. Cui. iForest: Interpreting random forests via visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):407–416, 2019. ISSN 1941-0506. doi: 10.1109/TVCG.2018.2864475.

[305] Y. Zhao, Z. Sun, E. Tian, C. Hu, H. Zong, and F. Yang. A CNN model for herb identification based on part priority attention mechanism. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2565–2571, 2020. doi: 10.1109/SMC42975.2020.9283189.

[306] X. Zhong, M. Zhang, T. Tang, B. Adhikari, and Y. Ma. Advances in intelligent detection, monitoring, and control for preserving the quality of fresh fruits and vegetables in the supply chain. *Food Bioscience*, page 103350, 2023. doi: 10.1016/j.fbio.2023.103350.

[307] Y. Zhong, B. Huang, and C. Tang. Classification of cassava leaf disease based on a non-balanced dataset using transformer-embedded resnet. *Agriculture*, 12(9):1360, 2022. ISSN 2077-0472. doi: 10.3390/agriculture12091360.

[308] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba. Learning deep features for discriminative localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2921–2929, 2016.

[309] S.-Y. Zhou and C.-Y. Su. Efficient convolutional neural network for pest recognition - exquisitenet. In *2020 IEEE Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pages 216–219, 2020. doi: 10.1109/ECICE50847.2020.9301938.

[310] Y. Zhou and G. Hooker. Interpreting models via single tree approximation, 2016.

[311] Y. Zhou, W. Wu, H. Wang, X. Zhang, C. Yang, and H. Liu. Identification of soil texture classes under vegetation cover based on sentinel-2 data with SVM and SHAP techniques. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 15:3758–3770, 2022. ISSN 2151-1535. doi: 10.1109/JSTARS.2022.3164140.

[312] B. Zhu and M. Shoaran. Tree in tree: from decision trees to decision graphs. *Advances in Neural Information Processing Systems*, 34:13707–13718, 2021.