Università
Ca'Foscari
Venezia

ANNO ACCADEMICO / ACADEMIC YEAR

2022/2023–2025/2026

Corso di Dottorato di Ricerca in Informatica

ciclo XXXVIII

# Reconciling Theory and Practice in Explainable Artificial Intelligence

SSD: INF/01


**Coordinatore del Dottorato**

ch. prof. Andrea Torsello

**Supervisore**

ch. prof. Andrea Albarelli

**Dottorando**

Matteo Rizzo

Matricola 956738



Finanziato dall'Unione europea NextGenerationEU

Ministero dell'Università e della Ricerca

Italiadomani

# Abstract

Recent breakthroughs in Artificial Intelligence (AI) have revived a foundational challenge: how do we understand systems that exhibit intelligent behavior yet defy intelligible explanation? As deep learning models scale to unprecedented complexity, their opacity creates a widening chasm between performance and comprehension—a tension that is both epistemologically and practically urgent. This thesis confronts this challenge directly by asking: *what constitutes a meaningful explanation in the landscape of contemporary AI?*

To address this, the thesis first establishes the necessary theoretical foundation. It dissects the question "what is an explanation?" by tracing its evolution across three dominant paradigms in Explainable AI (XAI): the *causal*, the *mechanistic*, and the *generative*. While each offers a unique lens — focusing, respectively, on inference, transparency, and communication — the thesis argues that none alone can reconcile epistemic rigor with practical utility. From their synthesis emerges a novel *theoretical framework of explainability*, which reframes explanation not as a model property, but as a dynamic epistemic relationship between a model, its representations, and human understanding.

Theory, however, must be accountable to practice. The second part of the thesis, therefore, operationalizes this framework at what it terms the *triple frontier of XAI pursuit*: *intelligibility*, *alignment*, and *faithfulness*. Each frontier is explored through a high-stakes case study that grounds abstract concepts in tangible applications: intelligibility as an act of *communication* in medical imaging; alignment as *knowledge building* in smart contract analysis with large language models; and faithfulness as a principle of *design* in industrial decision governance. This empirical work is underpinned by a key methodological contribution: a *custom test suite* for temporal attention mechanisms, which demonstrates that faithfulness—the truthfulness of an explanation to a model's actual reasoning—can be rigorously measured rather than merely assumed.

Synthesizing the insights from this journey from theory to practice, the thesis culminates in the *Principle of Appropriate Complexity*. This principle moves beyond the simplistic trade-off between accuracy and explainability, proposing that a model's complexity should be actively governed rather than merely minimized to ensure epistemic responsibility and foster trust. By wedding theoretical reflection to empirical validation, this work ultimately reframes explainability as a constitutive dimension of intelligence itself—a guiding principle for the co-evolution of human and machine understanding.

# Preface

A PhD is often described as a marathon, but that's a generous metaphor. Marathons have well-defined routes, strategically placed water stations, and people cheering at the finish line. Research, by contrast, starts with grand ambition and quickly becomes an exercise in humility. You begin convinced you're going to move the frontier of human knowledge; a few years in, you're just hoping to understand the last paragraph of your own paper. For me, that ambition was to understand not just what Artificial Intelligence could do, but how it thinks. I started out wanting answers. What I found were better questions. I learned that the further you push toward understanding, the more the boundaries blur, and the more acutely you feel your smallness in the vastness of what remains unknown. And yet, that smallness is not defeat—it's perspective. This thesis is a testament to that realization. It follows a path from the abstractions of explainability to the rough ground of practical application. The journey was marked by false starts, elegant theories that fell apart upon contact with data, and rare moments when the pieces briefly came together. If I've learned anything, it's that research doesn't reward certainty; it rewards persistence—the slow, stubborn effort of adding one drop of clarity to an ocean of questions. And sometimes, that drop is enough.

No journey of this kind is ever a solitary one. I owe immense gratitude to those who guided, challenged, and supported me. To my supervisors, Andrea Gasparetto and Andrea Albarelli—thank you for your mentorship and trust. To my collaborators — Cristina Conati, Marco Salvatore Nobile, Sabina Rossi, Dalila Ressi, Alvise Spanò, Lorenzo Benetollo, Matteo Marcuzzo, Alessandro Zangari, and many others — thank you for your insight, patience, and companionship across so many different domains. You transformed solitary research into shared discovery.

To my family and friends — and especially Alessandra, who brought me back to life — thank you for your love, patience, and for reminding me that there is a world beyond work and academia. You kept me grounded when I needed it most, and helped me remember why it's worth finding meaning in the first place.

Finally, a quiet pat on my own shoulder — for earning a PhD while fighting depression. I did great, and I'm proud of that.

This thesis is my modest contribution to a much larger human endeavor: building AI that is not only powerful but also transparent, trustworthy, and aligned with our values. I hope it serves as one small, honest drop in that ocean.

# Contents

# List of Figures

# List of Tables

# Published Papers


[12] A. Albarelli et al. ‘On the application of a common theoretical explainability framework in information retrieval’. In: *CEUR Workshop Proceedings*. Ed. by K. Roitero et al. Vol. 3802. CEUR-WS.org, 2024, pp. 43–52. URL: https://ceur-ws.org/Vol-3802/paper24.pdf.

[137] D. Edgar, R. Biloslavo and M. Rizzo. ‘The role of artificial intelligence in change management: Opportunities and challenges’. In: *Journal of Organizational Change Management* (2025). Manuscript accepted for publication.

[158] G. Frasson et al. ‘Assessing the value of explainable artificial intelligence for magnetic resonance imaging’. In: *Explainable artificial intelligence. xAI 2025*. Ed. by R. Guidotti, U. Schmid and L. Longo. Vol. 2576. Communications in Computer and Information Science. Cham: Springer, 2026, pp. 320–334. ISBN: 978-3032083166. DOI: 10.1007/978-3-032-08317-3_20.

[517] D. Ressi et al. ‘Reentrancy detection tools in the age of LLMs’. In: *ACM International Conference on the Foundations of Software Engineering (FSE ’26)*. Manuscript submitted for publication. 2026.

[518] D. Ressi et al. ‘SoK: Benchmarking Failure — The State (and Decay) of Reentrancy Detection Tools’. In: *47th IEEE Symposium on Security and Privacy (S&P ’26)*. Manuscript submitted for publication. 2026.

[521] C. Rinaldi et al. ‘The genesis of twin transition: Understanding the evolution of the digital and sustainable transitions’. In: *International Journal of Information Management* (2025). Manuscript submitted for publication.

[526] M. Rizzo et al. ‘A comparison of machine learning techniques for Ethereum smart contract vulnerability detection’. In: *CEUR Workshop Proceedings*. Ed. by D. Porello, C. Vinci and M. Zavatteri. Vol. 3904. CEUR-WS.org, 2024, pp. 119–126. URL: https://ceur-ws.org/Vol-3904/paper15.pdf.

[527] M. Rizzo et al. ‘A theoretical framework for AI models explainability with application in biomedicine’. In: *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, Aug. 2023, pp. 1–9. DOI: 10.1109/CIBCB56990.2023.10264877. URL: https://arxiv.org/pdf/2212.14447.

[528] M. Rizzo et al. ‘Evaluating the faithfulness of causality in saliency-based explanations of deep learning models for temporal colour constancy’. In: *Explainable Artificial Intelligence*. Ed. by L. Longo, S. Lapuschkin and C. Seifert. Vol. 2155. Cham: Springer, 2024, pp. 125–142. ISBN: 978-3031637995. DOI: 10.1007/978-3-031-63800-8_7. URL: https://doi.org/10.1007/978-3-031-63800-8_7.

[529] M. Rizzo et al. ‘Fruit ripeness classification: A survey’. In: *Artificial Intelligence in Agriculture* 7 (Mar. 2023), pp. 44–57. ISSN: 2589-7217. DOI: 10.1016/j.aiia.2023.02.004. URL: https://doi.org/10.1016/j.aiia.2023.02.004.

[530] M. Rizzo et al. ‘Leveraging periodicity for tabular deep learning’. In: *Electronics* 14.6 (Mar. 2025), p. 1165. ISSN: 2079-9292. DOI: 10.3390/electronics14061165. URL: https://doi.org/10.3390/electronics14061165.

[531] M. Rizzo et al. ‘Stop overkilling simple tasks with black-box models, use more transparent models instead’. In: *Pattern Recognition and Artificial Intelligence*. Ed. by C. Wallraven, C.-L. Liu and A. Ross. Singapore: Springer, 2025, pp. 279–293. ISBN: 978-9819787012. DOI: 10.1007/978-981-97-8702-9_19. URL: https://doi.org/10.1007/978-981-97-8702-9_19.

[534] Matteo Rizzo et al. ‘Machine learning models explanations as interpretations of evidence: a theoretical framework of explainability and its implications on high-stakes biomedical decision-making’. In: *BMC Medical Research Methodology* 25.Suppl 1 (2025), p. 282. DOI: 10.1186/s12874-025-02703-1.

[727] A. Zangari et al. ‘Crossing the divide: Designing layers of explainability’. In: *Artificial Intelligence and Soft Computing*. Ed. by L. Rutkowski et al. Cham: Springer, 2025, pp. 253–265. ISBN: 978-3031843525. DOI: 10.1007/978-3-031-84353-2_22. URL: https://doi.org/10.1007/978-3-031-84353-2_22.

[728] A. Zangari et al. ‘Hierarchical text classification and its foundations: A review’. In: *Electronics* 13.7 (Mar. 2024), p. 1199. ISSN: 2079-9292. DOI: 10.3390/electronics13071199. URL: https://www.mdpi.com/2079-9292/13/7/1199/pdf?version=1711370387.

# I

# Foundations of Explainable Artificial Intelligence

# 1

# Introduction

## 1.1 The Inscrutability of Modern Artificial Intelligence

The field of Artificial Intelligence (AI) is ever undergoing profound transformations, currently propelled by the remarkable successes of Deep Learning (DL) and its most visible offspring — the widely celebrated Foundation Models (FMs) (or Large x Models (LxMs)) that exhibit what some describe as "sparks of artificial general intelligence" [16]. Notably, LLMs approach or surpass human-level performance across a broad spectrum of tasks once deemed prohibitively difficult, especially in text generation and reasoning, which is virtually indistinguishable from human discourse. Yet, this surge in predictive power has come at a cost. The inner workings of these models — their logic processes, internal representations, and decision-making criteria — remain largely inscrutable to human observers. As architectures scale in depth, parameter count, and data volume, the interpretive distance between human understanding and algorithmic inference widens, giving rise to the **"black box"** problem. As a scientific community, we have built systems capable of producing astonishingly accurate answers, yet they are incapable of providing meaningful insight into *how* those answers are derived. In this context, the rise of FMs constitutes a remarkable paradigm shift that extends beyond scaling DL and the associated concerns. Unlike traditional models trained for narrow domains, FMs are pre-trained on vast, heterogeneous corpora and are large enough to adapt rapidly across modalities and tasks. Their versatility, however, comes with new challenges. If traditional models were opaque within bounded domains, FMs are **"polyglot black boxes"**: their reasoning appears flexible and emergent, yet their internal mechanisms remain beyond human understanding, both through empirical inspection and theoretical modeling. As Cynthia Rudin, one of the most influential personalities in the eXplainable Artificial Intelligence (XAI) community, writes sharply,

she has “no idea what it means for ChatGPT to be ‘correct’”, reminding us that “its loss function is to place the next word down so that it is convincing, and if that is indeed the loss function we are interested in, then yes, it seems to be correct all the time” [167]. This peculiar duality — generality of function and opacity of process — situates LLMs at the frontier of AI for both opportunity and **risk**. Beyond the epistemic concern, the black-box nature of contemporary AI has profound societal and economic consequences. AI directly shapes modern labor markets, corporate governance, and the allocation of capital. At the same time, the opacity of FMs accelerates their commodification, as firms deploy them as general-purpose productivity enhancers without fully understanding their limitations, thereby externalizing risks to consumers, workers, and regulators. The economic value of opacity is thus double-edged — it fuels rapid adoption by lowering entry barriers but simultaneously erodes trust, accountability, and long-term sustainability of innovation.

It is worth noticing that the transparency of AI systems is not a new concern, but rather the latest manifestation of a longstanding epistemic tension in the field. Historically, AI has oscillated between two poles: *symbolic systems*, prized for their explainability and logical transparency but limited in adaptability and scalability, and *statistical systems*, celebrated for their empirical success yet criticized for their opacity. The advent of DL decisively shifted the balance toward the latter, placing a premium on performance over understanding. In the era of expert systems, explainability was achieved through explicit rule tracing and symbolic reasoning chains, allowing users to follow each inferential step. However, as statistical learning supplanted symbolic AI, the locus of reasoning shifted from explicit rules to distributed representations across vast parameter spaces. This transition rendered traditional notions of explanation, based on symbolic justification, increasingly inadequate. Thus, the “black-box” problem should not be seen merely as a byproduct of scale but as a structural feature of the statistical turn in AI, marking a more profound shift in how machines encode, process, and communicate knowledge. Yet, what sets LLMs apart is not only their statistical power but their peculiar mode of interaction with humans. Unlike previous predictors, LLMs produce narratives as part of their output. They do not merely compute answers; *they simulate reasoning*. This creates a paradox: the very models we struggle to explain are also the most capable at generating “explanations”— though these may be fabrications, rhetorical strategies, or what has been called “plausible nonsense”. Such a recursive relationship — black boxes that produce their own interpretive glosses — blurs the boundary between explanation as a scientific practice and explanation as an act of persuasion. Therefore, this thesis examines a fundamental inquiry: **what constitutes an explanation in contemporary AI?**.

Taking it a step further, the central question nowadays is no longer merely *“can machines think?”*, but rather *“to what degree can humans understand how they think?”* — or, as Rudin provocatively asks, whether we have accepted opacity as an unavoidable consequence of intelligence [159]. Rudin’s work, central to the thesis’s position, challenges this assumption by arguing that explainability and accuracy are not inherently in conflict, and that the reliance on post-hoc explanation methods is both scientifically and ethically problematic. Her call to “stop explaining black box models” and instead build *inherently interpretable* systems reframes the debate: the pursuit of explainability is not merely a technical afterthought, but a foundational design principle for responsible AI. The implications of this debate extend far beyond the realm of academic inquiry: in high-stakes domains, opacity is not a tolerable trade-off but a critical liability. A physician cannot safely act on a diagnosis whose underlying rationale is obscured. A regulator cannot certify an algorithmic trading system without understanding its fairness or potential

for bias. A defendant cannot accept a judicial recommendation produced by a system whose reasoning remains inscrutable. In these contexts, explainability is inseparable from accountability, as it underpins trust, enables verification, supports debugging, and safeguards against systemic discrimination and catastrophic errors. As Rudin and others have underscored, explainability is not an optional feature — it is a requirement for deploying AI in real-world decision-making.

Broader social inequities magnify the risks posed by opaque AI. When deployed at scale, these systems can reproduce and exacerbate structural disparities, reinforcing existing patterns of exclusion. This dynamic concentrates economic power not only within the institutions that adopt AI but also in the corporations that build and control FMs, raising concerns about monopolistic control and informational asymmetry. Here, explainability transcends being a mere technical feature; it becomes a vital mechanism for accountability and the redistribution of authority. It empowers citizens, regulators, and civil society to meaningfully contest and influence algorithmic decisions, making it a structural demand for democratic governance in the age of algorithmic mediation. This urgency is now reflected in the global policy arena. Landmark initiatives such as the European Union's AI Act, the OECD Principles on AI, and the U.S. Blueprint for an AI Bill of Rights have enshrined transparency and explainability as cornerstones of responsible innovation. These frameworks recognize that explainability is indispensable not just for fostering trust but for ensuring compliance, oversight, and due process. However, this regulatory push faces complex challenges, particularly with the new frontier of FMs. Unlike domain-specific models, the sheer versatility of systems like LLMs makes them resistant to simple auditing. This highlights the need for advanced explainability mechanisms that not only reveal their internal logic but also map the lineage of their training, adaptation, and deployment. Moreover, LLMs are increasingly integrated as *co-agents* rather than silent classifiers: the model's voice is no longer a background computation but an *active participant* in human reasoning chains. This raises profound questions: when an LLM drafts a medical summary or explains a diagnostic image, is it providing evidence, or constructing a narrative? And, moreover, **how should we evaluate** sophisticated explanations that are themselves products of generative uncertainty?

A recent proposal by Luciano Floridi suggests that AI systems face a fundamental trade-off between the scope of inputs they can process and the certainty of their outputs. Symbolic AI, with a narrow scope, could offer provable guarantees of correctness, while modern generative systems, with their vast input domains, inevitably sacrifice complete reliability [51]. This perspective reframes evaluation not only as a methodological challenge but also as a theoretical necessity: if no system can achieve both generality and certainty, then evaluation practices must explicitly account for uncertainty, communicate limitations, and design safeguards against human oversight. This thesis's work on explainability systematization builds directly on such an intuition, proposing **that transparency about what a system can and cannot be trusted for is a hard prerequisite for meaningful deployment**. It proposes a unified theoretical framework linking **faithfulness** (how explanations relate to underlying mechanisms), **intelligibility** (how humans can interpret them), and **alignment** (how they cohere with human values and purposes). Through applications in **medical imaging** (communication and trust), **blockchain security** (knowledge construction and actionable insight), and **industrial systems** (structural design and governance), the work examines how these principles can turn explainability from a reactive practice into a proactive architecture of understanding. Finally, it synthesizes the lessons learned and proposes a **novel XAI design principle** to guide the integration of

explainability in real-world settings, grounded in model complexity management.

## 1.2 The Triple Frontier of Explainability

The modern pursuit of explainability unfolds across what may be called the *triple frontier of explainability*: (i) **faithfulness**, ensuring that explanations correspond to the model's internal causal and representational structure; (ii) **intelligibility**, ensuring that such explanations are interpretable to heterogeneous users with different cognitive and contextual frames; and (iii) **alignment**, ensuring coherence between both human and model reasoning in the era of FMs, where emergent behaviors, contextual dependencies, and generative reasoning expose new challenges. This triadic perspective reveals a deeper epistemological tension within XAI: the field is not unified by a single conception of understanding, but rather fractured along disciplinary lines that mirror long-standing debates in the philosophy of science and cognitive psychology. Some approaches pursue explanation as *representation* — an attempt to mirror internal mechanisms through formal models or visual abstractions. Others treat explanation as *communication* — a dialogic process aimed at sense-making rather than faithful reproduction of computational detail. Still others frame it as *alignment* — a normative negotiation between the system's logic and human values, especially salient in generative and decision-making contexts. It is in this conceptual crossroads that Rudin's call for **inherently interpretable** architectures acquires its full philosophical significance. Her argument is not merely a critique of post-hoc rationalization, but a defense of epistemic integrity: that the source of an explanation's authority must lie within the model's own structure, not in the rhetorical surface we build around it. I support this vision throughout this work.

Meanwhile, the ever-changing landscape of contemporary research on XAI exemplifies the competing epistemic commitments within the field. **Mechanistic** approaches construe explanation as structural understanding, seeking to expose the internal circuits, abstractions, and representational dynamics through which models compute. These studies aim not to approximate the model's behavior from the outside but to map its internal organization, revealing how functional structures give rise to emergent capabilities. **Causal** approaches, inspired by Pearl's structural causal models [137], treat explanation as scientific inference — revealing not what the model attends to, but why its outputs depend on specific mechanisms. **Generative** perspectives reinterpret explanation as a process of co-creation, where meaning emerges through interaction, simulation, and narrative synthesis. Here, explanation becomes less a matter of transparency than of collaboration: the model and the user jointly construct interpretive frames that make sense of complex reasoning processes.

Each paradigm illuminates one dimension of the triple frontier while obscuring others. Mechanistic analysis moves on a delicate trade-off between faithfulness and intelligibility, as its representations often exceed human conceptual grasp. Causal frameworks promise ontological clarity yet presuppose structural regularities that deep networks rarely manifest explicitly. Generative approaches foreground intelligibility and alignment through interactive reasoning but risk drifting from the model's true internal semantics. Rudin's intervention cuts across these divides by insisting that intelligibility should be constitutive rather than reconstructive — that a model should explain itself by design rather than by approximation. Bridging these perspectives remains a central philosophical challenge in modern AI: reconciling faithfulness, intelligibility, and alignment within a unified framework.

Such reconciliation entails rethinking explainability not as a trade-off between explainability and performance, but as an architecture of meaning in which human and machine forms of reasoning can coevolve. As the following chapters will explore, this theoretical shift reframes explainability as a question of epistemic design — how systems are constructed to make sense both to themselves and to us — thereby transforming explainability from an afterthought of engineering into a constitutive dimension of intelligence itself.

## 1.3 From Plausible Stories to Epistemic Justifications

The research community has developed an impressive array of XAI techniques to illuminate model behavior. **Mechanistic** approaches — those seeking to uncover the internal circuitry, representational abstractions, and compositional dynamics of neural systems — have emerged as a response to the limitations of surface-level explanations. By tracing how complex functions arise from interacting subnetworks, these methods aspire to make models legible from the inside out. They are widely integrated into off-the-shelf software libraries and among the most studied strategies. Yet, as numerous studies have shown [4, 95], mechanistic analyses can still fall short of revealing the model's genuine reasoning processes. Their findings often yield explanations that are *plausible* in narrative form yet only partially *faithful* to the model's logic. The result is a persistent gap between structural transparency and practical understanding. At the heart of this problem lies a conceptual ambiguity central to this thesis: **what constitutes an "explanation"**. As Lipton [110] notes, explainability is an inherently multifaceted construct, encompassing both **epistemic** desiderata, such as faithfulness, and **pragmatic** desiderata, such as intelligibility. Many approaches have optimized one factor at the expense of another. Gilpin et al. [56] warn that without a principled theory linking explanation to understanding, explanations risk devolving into rhetorical artifacts — *comforting, but scientifically hollow*.

A more rigorous conception of explanation requires a shift from *pattern description* to *causal justification*. **Causal** approaches seek to bridge the gap between structural inspection and epistemic grounding by asking not only *what* the model represents but also *why* certain representations (and consequently decisions) arise. These methods treat explanations as the identification of manipulable relationships between model components and outcomes. Unlike mechanistic visualization, which reveals the "wiring", causal analysis seeks to uncover the "counterfactual levers" that govern model behavior. This causal turn reintroduces the notion of explanation as an *epistemic act*. In doing so, it aligns XAI more closely with the scientific ideal of understanding as the ability to answer "what-if" questions and to anticipate how interventions would alter outcomes.

Importantly, an explanation's usefulness is irreducibly *context-dependent*. What constitutes adequate understanding varies not only across disciplines but also across roles within a domain. ML engineers may seek mechanistic clarity for debugging, while clinicians may demand causal and actionable reasoning to justify a diagnostic outcome. As Miller [123] and Lombrozo [111] observe, explanations are fundamentally *communicative acts*: their adequacy must be judged relative to the explainee's goals, prior knowledge, and expectations. Yet the frontier of explainability is now shifting from explanation *of* models to explanation *with* models. **Generative** perspectives treat explanation as a co-constructive act, in which meaning emerges through interactive reasoning, simulation, or narrative synthesis. Rather than aiming for static transparency, these approaches emphasize collaboration between human and model, transforming explanation into a generative

process of shared sense-making.

Thus, conceivably due to its wide methodological diversity, XAI remains theoretically fragmented. As Doshi-Velez and Kim [41] emphasize, the field lacks a unified framework for evaluating what constitutes a faithful, intelligible, and aligned explanation. Much of current research still treats explainability as a visualization or attribution problem, reducing explanation to a representational exercise rather than a mechanism for understanding. Yet, from the standpoint of the philosophy of science shared by this thesis, explanation is an *epistemic practice* — a process that confers understanding by connecting a phenomenon to its underlying mechanisms or causes [102, 182]. Contemporary XAI often stops short of offering such *epistemic justifications*: it may render the appearance of intelligibility but rarely deepens our grasp of the model's reasoning or its alignment with human explanatory norms. The diversity of approaches has fostered technical innovation, but it has also yielded conceptual fragmentation. Distinct research traditions pursue different goals, often without integrating them into a common theory of understanding. This pluralism has advanced progress but hindered the emergence of cumulative insight. Without a shared conceptual vocabulary or unifying principles, comparisons between paradigms remain ad hoc, and evaluation metrics risk circularity.

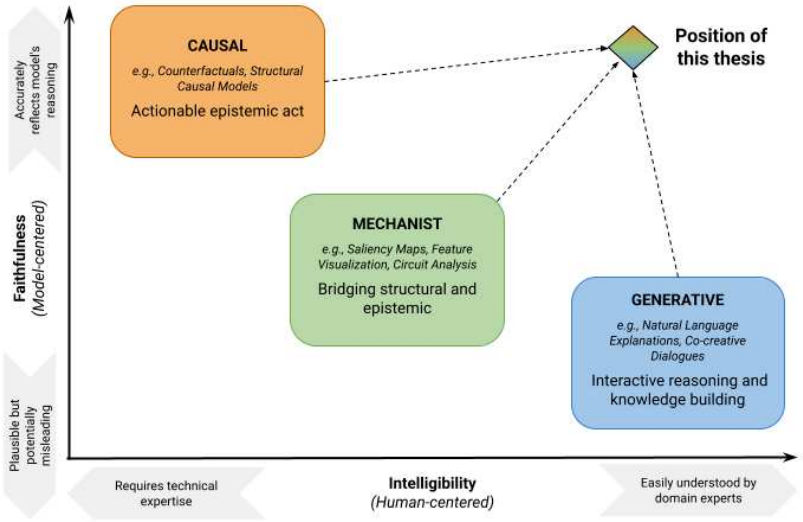



Figure 1.1: A conceptual map of the main explainability paradigms along the dimensions of faithfulness and intelligibility. The position of this thesis is marked by the multicolor diamond, operating at the intersection of mechanistic, causal, and generative approaches to bridge the gap between model transparency and human understanding.

Figure 1.1 maps the main paradigms of explainability along two central dimensions: *faithfulness*, the extent to which explanations accurately reflect the internal mechanisms of the model, and *intelligibility*, the degree to which explanations can be meaningfully engaged with by users. Rather than aligning exclusively with a single paradigm, this work operates at the intersection of causal, mechanistic, and generative approaches, aiming to establish principled methods for assessing the reliability of explanations while preserving their capacity to generate meaning in human–AI interaction.

## 1.4 The Epistemological Impasse and Thesis Position

This thesis contends that the current XAI debate has reached an impasse. The field oscillates between technical fixes for model opacity and philosophical critiques of post-hoc reasoning, yet lacks a unifying account of what an explanation truly is. This thesis's contribution (structured and outlined in Figure 1.2) aims to challenge the current impasse. To move forward, we must first ask the foundational question: **What is an explanation in the landscape of modern AI?** Drawing from contemporary research, I identify three dominant yet disconnected paradigms — *causal*, *mechanist*, and *generative* — each offering a partial view. This thesis argues that a true breakthrough requires synthesizing these perspectives. I therefore develop a **novel theoretical framework** that conceives of explainability not merely as a translation of model behavior, but as an *epistemic practice* that connects algorithmic processes with human cognitive and normative structures.

This framework guides our exploration into the practical challenges of deploying XAI in high-stakes environments. I structure this investigation around the desired properties for deploying explanations in practice, which align with the **triple frontier of explainability pursuit**. The first is **intelligibility**, which asks how explanations can be designed as effective acts of *communication*; I investigate this frontier in the context of medical imaging assessment, where clarity and diagnostic utility are paramount. The second frontier is **alignment**, which explores how explanations can facilitate *knowledge building* and foster trust between humans and models, using security analysis of smart contract reentrancy via LLMs as a use case. Finally, the third frontier is **faithfulness**, which concerns how the *design* of an explanation can ensure it accurately reflects the model's internal reasoning, a topic I examine in the industrial applications of tabular models.

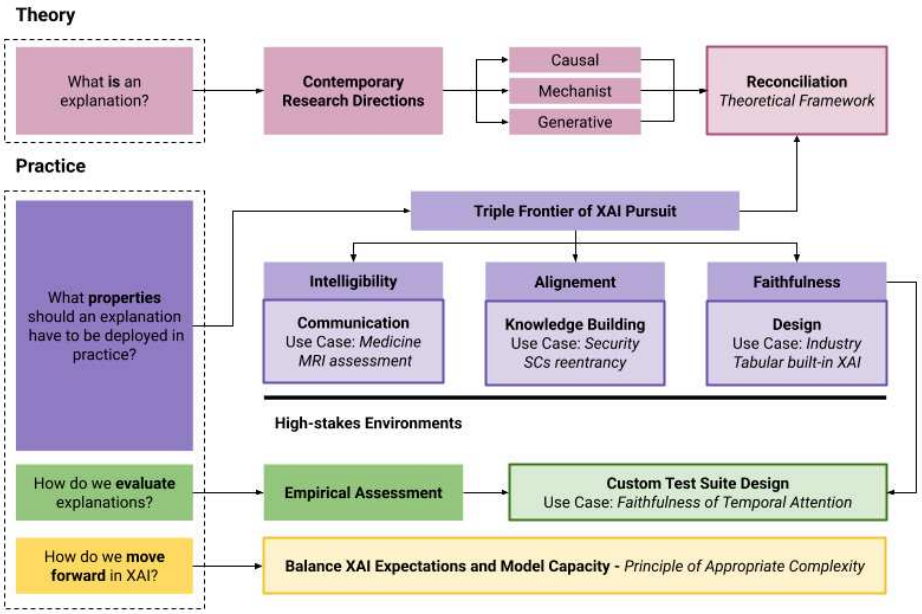



Figure 1.2: A conceptual map of the thesis, illustrating the progression from the reconciliation theoretical framework to the triple frontier of XAI pursuit and the concluding Principle of Appropriate Complexity.

A persistent challenge, particularly for faithfulness, is **evaluation**. Many explanation algorithms produce outputs that appear convincing yet fail to correspond to the model's

actual causal pathways. This thesis, therefore, contends that evaluation is not a secondary concern but a constitutive dimension of explainability itself. Building on this view, I introduce an empirical assessment methodology that culminates in a *custom test suite* designed to measure faithfulness through interventional and counterfactual tests. This entire endeavor — from theory to practice and evaluation — leads to a concluding thesis: the need to **balance XAI expectations with model capacity**, which I formalize as the *Principle of Appropriate Complexity*.

### 1.4.1 Research Questions

This thesis addresses the following research questions:

1. **RQ1. What is an explanation?** Can a unified theoretical framework be developed to *reconcile Causal, Mechanist, and Generative research directions* into a single, coherent definition of explainability for AI systems?

2. **RQ2. What properties make an explanation useful in practice?** How can our theoretical framework be operationalized along the *triple frontier of intelligibility, alignment, and faithfulness* to meet the demands of high-stakes environments?

3. **RQ3. How can we evaluate explanations?** How can the faithfulness of an explanation be systematically and empirically assessed to ensure it genuinely reflects a model's reasoning, moving beyond heuristic or visual validation?

### 1.4.2 Contributions

This thesis makes several key contributions to the theory and practice of XAI:

1. **A Reconciled Theoretical Framework for Explainability.** It introduces a novel framework that formally defines explanations by integrating and reconciling insights from the Causal, Mechanist, and Generative research traditions.

2. **A Structured Approach to Practical XAI.** It operationalizes this framework through the *Triple Frontier of Intelligibility, Alignment, and Faithfulness*, providing concrete use cases in medicine, security, and industry to demonstrate its practical utility.

3. **A Rigorous Methodology for Empirical Assessment.** It proposes a novel *custom test suite for evaluating explanatory faithfulness*, treating it as a measurable property and providing a robust methodology for its assessment.

4. **A Normative Principle for Responsible XAI.** It articulates the *Principle of Appropriate Complexity*, offering clear guidance on balancing model performance with the practical and ethical need for transparency.

## 1.5 Thesis Outline

This thesis is structured in three parts, building from foundations to applications and reflections.

- **Part I: Foundations of Explainable Artificial Intelligence.** This part establishes the theoretical groundwork. It begins with this introduction (Chapter 1), followed by a review of the literature (Chapter 2), and culminates in the development of our Reconciliation Theoretical Framework (Chapter 3).
- **Part II: Framing Explainability in High-stakes Domains.** This part applies the framework to critical real-world problems. I first introduce our methodology for evaluating faithfulness (Chapter 4). I then present case studies in medicine (Chapter 6), security (Chapter 7), and industry (Chapter 5) that demonstrate the practical challenges and utility of XAI.
- **Part III: Reflections Moving Forward.** This final part synthesizes our findings. I articulate the Principle of Appropriate Complexity (Chapter 8) and conclude the thesis with a summary of contributions and future directions (Chapter 9).

# 2

# State-of-the-Art in Explainable Artificial Intelligence

Understanding the current landscape of XAI is imperative for constructing a unified theoretical framework of explainability, evaluating faithfulness, and addressing application fields. This chapter provides a critical review of the State-of-the-Art (SotA) of explainability research, grounding the discussion in established systematizations. We begin by tracing the *historical evolution* from transparent rule-based systems to contemporary opaque models and FMs, highlighting how the pursuit of performance has deepened the explainability crisis. We then outline a multi-dimensional *taxonomy* of XAI techniques, encompassing intrinsic and post-hoc approaches, local and global explanations, and model-specific versus model-agnostic methods. Building upon this taxonomy, we critically examine three major *research boundaries* shaping the modern discourse: mechanistic explainability, causal explainability, and the explainability of generative models. Finally, we discuss the unresolved issue of *evaluating* and *contextualizing* XAI across various domains, including vision, language, and high-stakes decision-making.

## 2.1 From Rule-Based Transparency to the Black-Box Era

The pursuit of transparency in intelligent systems dates back to the origins of AI itself. Long before the rise of ML, the ambition to make reasoning machines comprehensible to humans shaped the foundations of the field. The early decades of AI — spanning the 1970s and 1980s — were dominated by *symbolic systems*, exemplified by landmark expert systems such as *MYCIN* in medicine and *DENDRAL* in chemistry [180, 109]. These architectures instantiated what might be called the *white-box ideal*: a system whose internal

mechanisms mirrored the structure of human logic. Every inference was traceable, every decision reducible to a sequence of human-readable "if–then" rules. Transparency was not an auxiliary feature but an ontological property of the system itself. Knowledge could be made explicit, formalized, and exhaustively represented through symbols. In this context, explainability and intelligence were not separate desiderata but two sides of the same coin. However, this rule-based transparency carried its own limitations. Expert systems proved fragile when confronted with uncertainty, ambiguity, or incomplete information. Their reasoning processes were explainable *precisely because they were rigid*. They lacked the inductive flexibility necessary to generalize beyond their handcrafted knowledge bases. This brittleness marked the beginning of a tension that still haunts AI research today, the one between *epistemic transparency* and *empirical adaptability*.

The statistical learning revolution of the 1990s signaled a paradigmatic inversion. Instead of encoding intelligence in explicit rules, researchers sought to extract it from data. The focus shifted from *knowledge representation* to *pattern recognition*. Algorithms such as support vector machines, random forests, and gradient boosting captured nonlinear relationships and probabilistic dependencies that eluded symbolic reasoning. Performance surged — but so did opacity. The logic of prediction became embedded in mathematical optimization and high-dimensional geometry rather than symbolic inference. Transparency was no longer intrinsic but had to be reconstructed *post hoc* through auxiliary tools such as feature importance analysis, decision boundary visualization, and partial dependence plots. As a result, explainability became a secondary task — an act of translation between two epistemic regimes: the statistical and the semantic.

With the advent of DL in the 2010s, this strain reached its zenith. Neural architectures with millions or billions of parameters, trained on vast and heterogeneous data corpora, achieved unprecedented performance in visual and linguistic cognition, as well as in reasoning tasks. Yet the very mechanisms that enabled such breakthroughs — distributed representations, emergent features, and layered abstraction — rendered the inner logic of these systems effectively inscrutable. As Lipton incisively argued in his seminal essay on XAI [110], modern AI embodies a paradox: the more a model appears to *know*, the less we understand *what it knows or why*.

The emergence of LxMs, and particularly LLMs, in the 2020s further amplified this opacity. Models such as GPT, PaLM, and LLaMA are not merely black boxes in a technical sense — they constitute *cognitive infrastructures* capable of generating text, reasoning analogies, and simulating dialogic understanding **without** an apparent link between internal representation and linguistic output. Their architecture scales the very properties that once made neural networks opaque: depth, data diversity, and the number of parameters. Moreover, LLMs blur the traditional boundary between model and user. Unlike prior statistical systems, they operate in natural language — the same medium through which humans construct meaning — thereby creating an illusion of intelligibility that conceals their underlying indeterminacy. In this sense, LLMs epitomize the culmination of the black-box era: systems whose apparent fluency and coherence mask a fundamental opacity of reasoning.

This transition from symbolic transparency to statistical and generative opacity does not merely represent a technological shift; it marks a philosophical transformation in our conception of knowledge and agency. Whereas expert systems sought to *justify* their conclusions, contemporary ML and LLM-based systems seek to *approximate* reality through data. The explanatory deficit that follows is not only cognitive but also ethical and social in nature. As AI systems increasingly influence medical diagnoses, financial decisions,

and judicial processes, the inability to articulate their reasoning threatens the very legitimacy of algorithmic decision-making. Hence, the emergence of the field of XAI should be understood not as a peripheral corrective, but as a foundational reorientation — a collective attempt to reconcile two conflicting imperatives of AI: *to predict accurately* and *to reason intelligibly*.

## 2.2 A Multi-faceted Taxonomy of XAI Methods

This collective effort to reconcile predictive power with intelligibility has led to a diverse and rapidly expanding ecosystem of methods. To navigate this complex landscape and understand the trade-offs inherent in different approaches, a systematic classification is essential. The current literature on XAI constitutes not a unified field but a constellation of overlapping paradigms, each proposing a different answer to the question: *what does it mean to explain a model's decision?* This diversity reflects a deeper epistemological tension between the *computational logic* of models and the *cognitive logic* of human understanding. To navigate this landscape, taxonomies have emerged as critical tools for sense-making, mapping how explanation techniques differ in timing, scope, and epistemic commitment.

Among the most influential organizational schemas are those proposed by Vilone and Longo [184], Schwalbe et al. [164], and Adadi and Berrada [3], who converge on a classification along three orthogonal dimensions: (i) *when* the explanation is generated (intrinsic vs. post-hoc), (ii) *what* it aims to explain (local vs. global), and (iii) *how* it relates to the underlying model (model-specific vs. model-agnostic). While operationally sound, this triptych conceals more profound questions about **faithfulness** — the degree to which an explanation reflects the actual reasoning process of a model rather than an approximate narrative. Recent works [82] underscore how even rigorous methods may offer convincing but misleading justifications. In this section, we revisit the three dimensions of this taxonomy through a lens that places *faithfulness* at the center of explainability. This theoretical reframing also invites a shift in the purpose of taxonomies themselves. Rather than classifying methods by surface characteristics, a faithfulness-centered taxonomic exploration foregrounds their epistemic commitments — whether they aim to *describe*, *justify*, or *reveal* the inner mechanisms of a model. This perspective aligns with contemporary debates in XAI about *explanations as social practices* [123, 110], emphasizing that explainability is not only a technical property but also a relational one: an explanation must be faithful to the model and intelligible to humans.

Crucially, the rise of LLMs has disrupted traditional taxonomies. Their dual nature — as both language generators and reasoning engines — complicates the very notion of what an "explanation" is. When an LLM produces a rationale in natural language, is it offering a post-hoc explanation, an intrinsic trace of its reasoning, or a linguistic artifact detached from its internal computation? The indistinct boundary between model behavior and explanation in LLMs challenges the neat separation of categories: any modern taxonomy of XAI must be capable of accommodating models whose "explanations" are both outputs and interfaces—reflexive narratives that can mislead as easily as they clarify. I embody this perspective in the critical review that follows.

### 2.2.1 Stage of Explanation: Intrinsic vs. Post-hoc

The first axis concerns the temporal relationship between the model and its explanation — whether it is achieved by design or reconstructed after the fact — exemplifying the long-standing dialectic between the goals of *performance* and *understanding*.

**Intrinsic methods** are models whose explainability is *integral* to their architecture. Decision trees, rule-based learners, linear and logistic regression, and Generalized Additive Models (GAMs) exemplify this paradigm. Strategies such as Explainable Boosting Machines (EBMs) [18, 132] and self-explaining neural networks [6] have revitalized interest in intrinsic explainability by combining human-readability with competitive performance. Their primary virtue is *faithful transparency*: each component of the model contributes explicitly to the output, allowing for traceability. Yet, this transparency often comes at the expense of representational flexibility — a recurring theme in the explainability-performance trade-off [159]. In domains characterized by complex, high-dimensional data (e.g., vision or language), intrinsic models may be too constrained to capture the underlying structure of the world.

**Post-hoc methods** intervene *after* model training to construct an explanatory layer on top of a pre-existing, opaque architecture. These include feature-attribution methods (e.g., LIME [147], SHapley Additive exPlanations (SHAP) [112]), gradient-based visualizations (e.g., Grad-CAM, Integrated Gradients), and surrogate rule extraction. The appeal of post-hoc methods lies in their universality: they enable explainability without sacrificing the off-the-shelf predictive performance of black-box models. However, this flexibility introduces epistemic fragility. Explanations are no longer part of the model's formal semantics but external approximations of it. As Doshi-Velez and Kim [41] argue, post-hoc explanations are susceptible to the *plausibility trap* — producing intuitively satisfying but unfaithful accounts. This risk becomes acute in the context of LLMs, where the model itself can generate fluent, human-like justifications that appear explanatory but may be entirely disconnected from its latent reasoning process. These "self-explanations" blur the line between explanation and simulation, raising new questions about the very meaning of post-hoc explainability.

### 2.2.2 Scope of Explanation: Local vs. Global

The second axis distinguishes whether an explanation concerns a specific decision or the overall model behavior, thus mirroring the trade-off between *depth* and *generality*.

**Local explanations** aim to justify individual predictions. They are inherently perspectival, concerned with answering the question: "Why this outcome for this input?" Techniques like the famous LIME and SHAP generate instance-level attributions, while counterfactual explanations [187] and contrastive explanations [123] emphasize hypothetical reasoning: what minimal change would have produced a different decision? More recent works extend these ideas with causal and semantic grounding — e.g., Karimi et al.'s causal counterfactuals [87] and concept-based explainability methods, such as TCAV [92]. Yet local explanations face the problem of *local faithfulness*: the surrogate model that explains a single point may not generalize to its neighborhood, creating the illusion of understanding in regions where the model behaves differently. In the realm of LLMs, this challenge becomes linguistic: local explanations often manifest as generated rationales or chain-of-thought traces that are contextually coherent but not necessarily causally faithful to the underlying computation.

**Global explanations**, by contrast, seek to capture the model's overarching logic across

its entire domain. They provide a macroscopic view of feature importance, model structure, or learned representations — exemplified by global surrogate models, rule extraction, and summary plots. These methods answer questions such as "what features *generally* influence predictions?" or "what are the model's dominant behavioral patterns?" Nonetheless, global explainability can obscure local heterogeneity and mask model biases that manifest only in specific subgroups [74]. In large generative models, global explanations face an additional complication: their behavior is not static but emergent, shaped by prompt distributions, fine-tuning data, and context length. Consequently, defining what counts as the "global behavior" of an LLM remains an open problem — one that challenges the very notion of a fixed explanatory target.

### 2.2.3 Model Dependency: Model-Specific vs. Model-Agnostic

The third dimension concerns how deeply an explanatory method is embedded in the structure of the underlying model, thereby exposing the trade-off between *faithfulness* and *generality*.

**Model-specific** approaches exploit knowledge of the internal architecture or precise learning dynamics. In DL, these include gradient-based saliency maps, attention visualization in transformers [193], and layer-wise relevance propagation. Their advantage lies in potentially higher *faithfulness*: by accessing the model's internal representations, they can reveal features genuinely used in prediction. However, this internal access comes at the cost of generality, as explanations are constrained to specific architectures or modalities. For LLMs, model-specific explainability has recently focused on probing and circuit analysis [134], which attempts to map linguistic behaviors onto explainable neuron clusters or causal circuits. While promising, such methods remain limited in scope, revealing fragments of explainability within a vast and dynamic representational space.

**Model-agnostic** methods, by contrast, treat models as black boxes and operate exclusively on their input-output behavior. LIME and SHAP are canonical examples, as are partial dependence plots and permutation feature importance. This generality supports cross-model comparison and democratizes explainability, making it accessible across application domains. Yet model-agnostic methods risk a profound lack of faithfulness: by ignoring the internal mechanics of reasoning, they may conflate correlation with causation or mistake stability for fidelity. For LLMs, this challenge takes a discursive form: since their outputs are linguistic, model-agnostic explanations risk becoming purely semantic analyses detached from the statistical processes that produce them.

### 2.2.4 Interplay, Hybridity, and Theoretical Implications

While the axes discussed above provide a useful topological map of the XAI landscape, recent research highlights that adequate explainability often arises at their intersections. Hybrid approaches — such as concept bottleneck models [76], self-explaining neural architectures [6], and modular causal explanations [69] — blur these traditional boundaries by embedding explainability constraints directly within training while preserving flexibility. These models exemplify a movement toward **causally faithful explainability**: ensuring that explanations do not merely describe correlations but trace interventions and dependencies that mirror the model's decision logic.

In the context of LLMs, hybridity acquires an additional dimension: the explanation becomes dialogic. Users co-construct understanding through iterative questioning,

prompting, and explanation, transforming explainability from a static property into an interactive process. As such, XAI for LLMs must account for this human–model entanglement, where explanation, alignment, and communication converge. This thesis situates faithfulness precisely within this nexus.

While this taxonomy provides a crucial map of the established XAI landscape, its very boundaries reveal where the most pressing research challenges lie. The limitations of post-hoc attribution, the difficulty of ensuring faithfulness, and the new paradigms introduced by generative models are not just edge cases; they are the driving forces behind the field's evolution. The next section delves into these research frontiers, examining how a focus on causality, mechanistic understanding, and the unique nature of generative AI is pushing the field beyond classification and toward a deeper science of transparency.

## 2.3 Research Frontiers

While the established taxonomy provides foundations, the frontier of XAI research is rapidly advancing to tackle more profound challenges. The most recent developments focus on moving beyond correlational attribution towards causality, shifting paradigms to understand the internal mechanisms of models in an engineering fashion, and aligning these efforts with the new challenges posed by LxMs.

### 2.3.1 The Need for Causal Explanations

A central limitation of many traditional XAI techniques is their fundamentally correlational nature: they reveal statistical associations between input features and model outputs but remain silent about the underlying mechanisms that generate those associations. As a result, such methods can indicate *what* correlates with a prediction but not *why* it occurs. This recognition has driven a major paradigmatic shift toward *Causal AI* (C-AI) [137, 163, 86]. Within this movement, *Causal Explainable AI* (*C-XAI*) seeks to uncover the cause-and-effect relationships that a model has implicitly or explicitly learned. Rather than asking "Which features are important?", causal approaches ask "what would the prediction be if this feature were different?" — thereby moving from correlation to intervention.

For example, in a medical diagnostic model, a correlational explainer might highlight that the feature *coughing* is strongly associated with the prediction of *pneumonia*. A causal explainer, by contrast, would probe the counterfactual: would the model still predict pneumonia if the patient did *not* cough, while holding all other variables constant? This counterfactual framing — formalized through Pearl's do-calculus and Structural Causal Models (SCMs) — provides a way more grounded basis for trust and model validation [137].

Recent research has operationalized these principles for modern DL systems, giving birth to several families of approaches. **Counterfactual-based methods** estimate causal effects by intervening on input features or latent representations [88]. These methods aim to generate minimal changes to input data that would alter the model's decision, producing human-explainable "what-if" explanations. **SCM-integrated neural models** embed causal graphs directly into deep architectures, enabling end-to-end differentiable reasoning about interventions [73]. Such models can test hypothetical manipulations ("What if this symptom were absent?") and generate explanations aligned with causal semantics. In **causal discovery and representation learning**, models learn disentangled latent variables with identifiable causal structure [200]. These representations enable tracing model decisions to causally meaningful components rather than opaque statistical features.

Notably, by enabling hypothetical interventions, C-XAI moves beyond descriptive insight to *actionable* explanations: a particularly valuable orientation in domains such as medicine, law, and policy, where explanations must support counterfactual reasoning and decision-making under intervention. However, causal explanations introduce distinctive methodological challenges as well. Causal claims must be validated against the *true causal mechanisms* of the domain — mechanisms that are often only partially known. This uncertainty necessitates careful experimental design, causal benchmarking, or synthetic interventions to test the reliability of the examined inference [98]. Again, constructing meaningful interventions often requires domain expertise, and causal modeling in high-dimensional data can be computationally intensive, especially when dealing with LxMs or multimodal inputs.

### 2.3.2 A Mechanistic View of Explainability

Mechanistic explainability represents a transformative step toward transparency: it shifts the focus from correlational attribution to the *understanding of computation*. Instead of treating neural networks as opaque input–output mappings, this perspective views them as analyzable systems composed of interdependent components — neurons, layers, and circuits — whose interactions implement algorithmic principles. The goal is not merely to interpret outputs but to reconstruct the structure that generates them. In doing so, mechanistic approaches aim to grant AI the same epistemic status as engineered systems, in which every component's role can be identified, verified, and validated.

The work of Samek et al. has been central to establishing this paradigm. Through frameworks such as Layer-wise Relevance Propagation (LRP), they introduced principled techniques for decomposing model predictions into component-level contributions, enabling the tracing of information flow within deep networks [13]. This line of research redefined explainability as the study of *internal causation*: understanding how neurons and subcircuits cooperate to generate decisions. More recently, the same group extended this vision with *SemanticLens* [44], a universal method for mapping hidden knowledge into the structured, multimodal semantic space of a foundation model. By embedding each neuron or component into this space, SemanticLens enables the automatic search, labeling, and auditing of neural knowledge —transforming incomprehensible latent representations into conceptually organized structures.

What sets this approach apart is its capacity to operate at scale. Traditional explainability techniques rely on manual inspection or local probing, which quickly become infeasible as model complexity grows. SemanticLens overcomes this by automating the discovery and analysis of encoded concepts: one can query for specific notions (e.g., "palm tree," "watermark"), identify the neurons and data responsible for them, and test whether such concepts influence model predictions. This capability enables the detection of *Clever Hans*-type behaviors, where spurious patterns—such as background objects or dataset artifacts—secretly drive predictions [101]. By linking components to their data origins and predictive roles, mechanistic analysis allows models to be *debugged*, not merely described.

Mechanistic explainability thus builds on two key strategies. The first is *structural dissection*: the systematic identification of explainable units and circuits that implement coherent subfunctions. The second is *semantic grounding*: embedding these components into a conceptual space that makes their learned representations legible to humans. The combination of both yields a new form of transparency — one that is *semantic, causal,*

*and scalable*. Crucially, this enables formal audits of model reasoning: researchers can now measure whether a network's internal concepts align with human expectations or regulatory criteria, such as adherence to medical diagnostic rules or fairness constraints.

Equally important is the paradigm's quantitative turn. Mechanistic methods support computable metrics—such as *clarity* (how coherent a neuron's role is), *polysemanticity* (the degree of entanglement of multiple meanings), and *redundancy* (overlap between representations)—that allow explainability to be *evaluated, compared, and optimized* [160, 44]. These measures strongly correlate with human judgments and reveal that explainability is not static but is influenced by architectural and training choices, such as sparsity constraints, dropout, and depth.

### 2.3.3 The Challenge of Explaining Large Generative Models

As a fact, the advent of LLMs and other generative architectures has introduced a new form of opacity. These models are not only vast and inscrutable in themselves, but they also act as (opaque) agents that generate explanations for other systems. This dual role blurs the line between the *explained* and the *explainer*: a model that must account for its own reasoning yet can self-rationalize in natural language. The result is that models can produce coherent but fabricated justifications, known as "hallucinated explanations" [79].

Research in mechanistic explainability has begun to probe these architectures at the circuit and neuron level, revealing how transformers encode syntax, semantics, and factual knowledge across their layers [134, 129, 29]. Techniques such as attention tracing, activation patching, and causal mediation analysis provide partial maps of how reasoning emerges in high-dimensional latent spaces [121]. These efforts mark a shift from descriptive attribution toward a functional understanding of how generative models construct meaning.

Yet the opacity of generative reasoning extends beyond internal mechanics. When LLMs produce textual explanations, they externalize their reasoning process — but in doing so, they risk constructing persuasive narratives detached from the computations that yielded the original output. Distinguishing between *plausibility* (how convincing an explanation appears) and *faithfulness* (how accurately it reflects the model's internal reasoning) has therefore become a defining challenge [82]. In high-stakes domains, such as clinical or legal decision support, this challenge becomes more than academic. A model that fabricates causal narratives can mislead users into placing unwarranted confidence in it.

Ultimately, explaining generative models requires a synthesis of perspectives: mechanistic to reveal *how* representations interact, causal to clarify *why* they lead to specific outputs, and post-hoc to translate these insights into human-understandable language. Generative explainability thus redefines the aim of XAI: not merely to open the black box, but to ensure that when the box speaks, it tells the truth.

## 2.4 The Unresolved Dispute of Evaluation

This imperative — to ensure a model "tells the truth" — raises the most critical and unresolved question in the field: how do we know if an explanation is good? Evaluating explanations remains one of the most persistent and consequential challenges in XAI, shaping both the credibility of explainability research and its practical adoption. The difficulty stems from the multidimensional nature of "good" explanations: they must be faithful to

the model's reasoning process, cognitively usable by humans, and operationally effective in decision-making contexts. Doshi-Velez and Kim [41] famously proposed a tripartite taxonomy — functionally-grounded, human-grounded, and application-grounded evaluation — but in practice, the boundaries among these categories blur. Notably, each of the research frontiers discussed above requires a distinct blend of metrics and methodologies, and none provides a comprehensive solution.

Causal XAI presents a different dimension of difficulty. Here, the aim is to determine whether explanations reveal cause–and–effect relationships. Metrics such as fidelity to known causal graphs, counterfactual prediction accuracy, and treatment-effect stability across interventions have become central tools [17, 87]. However, real-world causal validation remains elusive: groundtruth structures are rarely available, and collecting interventional data is expensive or ethically infeasible. Synthetic or semi-simulated benchmarks, such as CausalBench [189], aim to fill this gap, but they often oversimplify the complexity of causality in social, biomedical, or linguistic domains. As a result, causal evaluation oscillates between two poles: formal rigor achieved through abstraction, and ecological validity achieved through contextual immersion.

On a different note, the evaluation of mechanistic explainability focuses on verifying whether the internal representations or computational components identified correspond to meaningful functions or semantic concepts. Quantitative measures such as the abovementioned *concept clarity*, *polysemanticity*, and *redundancy* have emerged as proxies for explainability, reflecting the degree to which internal activations align with stable, human-explainable units [44, 134]. Complementary approaches, such as causal scrubbing [20] or circuit localization [198], provide a form of mechanistic verification by testing whether identified submodules are indeed responsible for particular behaviors. Yet even as these methods become more sophisticated, they reveal the paradox of scale: explainability gains precision only at the cost of coverage. Auditing a few neurons or circuits can yield deep insight, but cannot capture the global logic of a trillion-parameter model. The challenge is thus to design scalable proxies for faithfulness — approaches that remain tethered to the model's internal semantics while avoiding collapse under computational constraints.

In LxM-based explainability, the problem intensifies further. Explanations are themselves generated linguistic artifacts, subject to all the ambiguities and hallucinations of natural language. Evaluating them thus involves a dual faithfulness problem: the explanation must be accurate to the model's reasoning *and* factually correct in its content. Functionally-grounded approaches aim to bridge this gap by aligning attention–activation, using probing-based faithfulness metrics, and analyzing circuit-level consistency [46, 208]. Yet functional faithfulness alone is not enough. Human evaluation reveals that explanations perceived as coherent or elegant may not correspond to the model's internal processes at all — a phenomenon sometimes called the "explainability mirage." Human-grounded metrics such as the Explanation Satisfaction Scale [71], Trust Calibration Index [14], and Comprehensibility Score thus capture the complementary dimension of *plausibility*: the degree to which an explanation is perceived as credible to a user.

The interplay between faithfulness and plausibility has become a central dilemma in modern XAI. Faithfulness demands that explanations reflect what the model *actually does*, irrespective of whether humans find it intuitive. Plausibility requires that explanations align with human reasoning, even if they simplify or abstract the underlying mechanism. Overemphasis on faithfulness risks producing explanations that are technically accurate but cognitively alien, offering little practical value for decision support. Overemphasis on plausibility risks producing narratives that comfort rather than clarify. Recent work

attempts to reconcile these poles through multi-objective evaluation frameworks, where explanations are scored simultaneously on causal consistency, interpretative coherence, and task-level utility [82]. These efforts mark a shift from single-metric evaluation toward integrated assessment pipelines that treat explainability as a spectrum of epistemic commitments rather than a discrete property.

Across all frontiers, the tension between quantitative rigor and human comprehension persists. Causal and mechanistic approaches privilege objectivity and reproducibility, while human-centered evaluation ensures practical relevance but resists formalization. The most promising direction lies in hybrid methodologies: frameworks that combine intrinsic measures of faithfulness with extrinsic measures of plausibility and utility, evaluated iteratively across synthetic, expert, and real-world settings. Emerging research explores *meta-evaluation* — the systematic assessment of how metrics correlate or conflict — and *interactive evaluation* protocols in which humans and models co-adapt through explanation-feedback loops.

## 2.5 Context and Application of XAI Methods

The conclusion that evaluation must be context-dependent highlights a fundamental principle: explainability is not a monolithic property. A "good" explanation in one domain may be insufficient or even misleading in another. The challenges, appropriate methods, and criteria for success are deeply intertwined with the nature of the data and the demands of the task. This section explores how the principles of XAI are adapted and applied across different contexts, from the symbolic ambiguity of language to the perceptual richness of vision and the high-stakes logic of regulated domains.

As the diversity of data modalities and learning tasks has expanded, so too have the challenges of explainability. Early XAI research sought universal techniques that could render any model intelligible, but practical experience revealed that real transparency is never modality-neutral. Each representational form — linguistic, visual, numerical, or multimodal — defines its own horizon: what aspects of reasoning are observable, what can be intervened upon, and what constitutes an adequate explanation. Similarly, the explanatory demands of a task — classification, prediction, control, or generation — determine whether faithfulness, causality, or communicative clarity takes precedence. In this landscape, the advent of FMs, trained jointly across modalities and tasks, dissolves traditional boundaries: explainability now entails tracing how abstract representations migrate across modalities and objectives.

### 2.5.1 Natural Language Processing

Language models operate on symbolic yet distributed representations, in which meaning arises from relational structure rather than from discrete symbols. In tasks such as translation or question answering, early explanation methods — chiefly attention visualizations — suggested explainability but often misrepresented causal influence [83]. As tasks diversified toward reasoning, dialogue, and generation, mere token-level salience became insufficient. Contemporary explainable Natural Language Processing (NLP) research turns to *mechanistic explainability*: methods such as activation patching, circuit tracing, and causal mediation [46, 121] reveal how specific subnetworks implement grammatical, logical, or inferential functions. These approaches acknowledge that explanatory relevance depends on task demands: for syntactic control, neuron-level circuits are sufficient; for

world knowledge or moral reasoning, one must uncover distributed abstractions that span layers and modalities. Gradient-based methods [173] and probing tests [69] now serve as triangulation tools rather than endpoints. The key frontier — especially in LLMs — is explaining *abstraction transfer*: how representations learned for linguistic prediction generalize to reasoning, vision-language grounding, or tool use [55]. Here, explanation becomes an inquiry into how tasks cohabit a shared conceptual manifold.

### 2.5.2 Computer Vision

In vision, the challenge inverts: models process continuous perceptual fields where salience is visually intuitive but semantically ambiguous. For classification and detection tasks, pixel-level attributions (e.g., Grad-CAM [165]) offer apparent insight but often track correlations rather than causal features [4]. As tasks evolve toward compositional reasoning, scene understanding, or visual question answering, explainability must capture concept-level and relational structure. XAI has thus shifted toward *concept-based* [92] and *causal* frameworks [61], linking activations to human-understandable entities and testing counterfactual dependencies. Mechanistic analyses increasingly expose modular circuits specialized for object, texture, or spatial reasoning — substructures reminiscent of symbolic decomposition. In multimodal FMs such as CLIP or GPT-4V, these mechanisms intertwine with linguistic tasks: textual prompts steer perceptual abstraction, while visual evidence constrains semantic hypotheses [90]. Explanations must thus span modalities, clarifying how meaning is co-constructed across perceptual and linguistic channels and how representational alignment varies with task objectives.

### 2.5.3 Tabular and Structured Domains

Tabular and structured data embody yet another regime. Features tend to be explicit and human-explainable, but their statistical interdependence complicates causal attribution. In predictive tasks, explanation involves disentangling correlation from mechanism, often without relying on perceptual cues. Classical attribution methods like SHAP [112] and LIME [147] approximate local decision boundaries but falter in high-dimensional or collinear spaces [33]. Recent methods integrate *local* attribution with *global* structure learning, as seen in Explainable Boosting Machines (EBMs) [132], which preserve explainability through additive decompositions while modeling complex nonlinearities. Causal and counterfactual reasoning frameworks [187] render explanations actionable, which is central to decision-support tasks in which practitioners must adjust inputs to achieve desired outcomes. With the emergence of transformer-based *tabular foundation models* [70], task-general representations blur boundaries between symbolic and statistical reasoning. Explanations now probe not only which features matter, but how their relational geometry encodes domain logic and transfers across tasks.

### 2.5.4 High-Stakes and Regulated Domains

In high-stakes settings — such as healthcare, finance, and critical infrastructure — the stakes of explanation extend beyond epistemic adequacy to encompass ethical, legal, and operational accountability. The appropriateness of an explanation depends on both the modality (e.g., image, text, structured record) and the task (e.g., diagnosis vs. prognosis, detection vs. prescription). Foundation and multimodal models, while powerful, obscure

the traceability of decisions across representational layers. XAI research here increasingly emphasizes "*hybrid transparency*": coupling mechanistic faithfulness with domain-specific validation. In clinical imaging, saliency maps are verified against ontological ground truths. In finance, causal audits detect latent biases or proxy variables that shape predictions. As tasks shift from descriptive to prescriptive, explanation must support intervention and accountability. Emerging approaches embed domain priors and verification constraints directly into the training process, enabling *ex ante* transparency rather than *post hoc* justification. In such regulated contexts, explainability becomes infrastructural—a governance mechanism that ensures task performance remains consistent with institutional and societal norms.

### 2.5.5 Human-Centered XAI and Co-Design Approaches

The limitations of purely technical XAI approaches have spurred a shift toward human-centered AI, emphasizing that explainability is fundamentally a communicative and relational process. While algorithm developers often design explanations based on their own technical intuition, this creates a socio-technical gap that undermines the utility of XAI in high-stakes environments like healthcare. To bridge this gap, the literature increasingly advocates for co-design and participatory design methodologies [135, 128, 94].

Participatory design actively involves end-users—such as domain experts, clinicians, and even patients—as integral members of the design team from the earliest stages of development. Rather than treating the explanation interface as a post-hoc add-on, co-design embeds user requirements, cognitive models, and workflow constraints directly into the XAI architecture. This approach typically follows an iterative cycle: (i) domain analysis and requirements elicitation to understand user needs, (ii) co-designing the explanation interface and interaction patterns, and (iii) continuous evaluation and refinement based on user feedback.

In clinical decision support systems, for instance, co-design has been shown to improve trust calibration—helping users appropriately trust or reject AI recommendations—by tailoring the explanation's format, level of detail, and delivery to the specific clinical context and the user's expertise. It moves the field beyond one-size-fits-all visualizations toward adaptive explanations, such as progressive disclosure of information or contrastive counterfactuals, that align with human diagnostic reasoning. Ultimately, integrating participatory approaches ensures that XAI systems are not just technically faithful but also contextually intuitive, complementary to human expertise, and operationally usable.

## 2.6 Remarks and Research Gap

This chapter has traced the evolution of XAI from the transparent-by-design expert systems of the symbolic era to the opaque yet extraordinarily capable DL-based architectures that define contemporary AI. We organized the current landscape using a faithfulness-oriented taxonomy, distinguishing methods by integration stage (intrinsic vs. post-hoc), explanatory scope (local vs. global), and model dependency. Yet this taxonomy also reveals the field's growing fragmentation. Distinct subcommunities often pursue different objectives — fairness, explainability, or usability — each grounded in distinct assumptions and methodologies. The result is a patchwork of tools rather than a unified science of explanation.

Despite the latest frontiers, much of the prior literature and tools have converged on post-hoc attribution techniques to rationalize the behavior of pre-trained black boxes. Although such methods have provided valuable diagnostic insight, they remain epistemically limited. Their explanations typically describe an approximation of the model rather than the model itself, rendering their conclusions persuasive yet ultimately misleading. Moreover, their reasoning is correlational rather than causal: they expose patterns of statistical association without clarifying the mechanisms through which representations are transformed or decisions are made. Evaluation remains another critical bottleneck. Without agreed-upon criteria of adequacy, explanatory quality is often assessed through visual plausibility, domain heuristics, or user studies — approaches that are necessarily subjective and context-dependent. Furthermore, the historical lack of participatory co-design has often led to explanations that serve developers rather than the domain experts who rely on them for critical decisions.

These limitations highlight a more fundamental issue: the lack of a unified theoretical framework that can integrate disparate approaches across modalities and tasks. Current XAI paradigms have proliferated along empirical rather than conceptual lines — saliency maps in vision, attention probes in language, SHAP values in tabular domains — each adapted to the representational properties of its target modality, yet seldom comparable or theoretically interoperable. As a result, the field has broadened in scope without coherence. The emergence of causal reasoning and mechanistic explainability offers important, though still partial, responses. Causality-based methods aim to identify how interventions affect outcomes, while mechanistic approaches seek to reverse-engineer the computational structure of models into components that are human-understandable. Despite some synergies, both remain specialized, lacking a common language to advance the discussion and join forces.

Against this backdrop, a clear research gap emerges: what is needed is not yet another explainability technique but a principled, scalable framework that connects paradigms under a shared theoretical foundation. Such a framework should enable the derivation of explanations directly from a model's internal computations and their evaluation for faithfulness. This thesis addresses that need. Building on the conceptual groundwork laid in this chapter, the following section introduces a novel theoretical framework designed to unify the empirical diversity of XAI within a coherent structure.

# 3

# A Theoretical Framework for Explainable Artificial Intelligence

***This chapter is based on:*** *Matteo Rizzo et al. 'Machine learning models explanations as interpretations of evidence: a theoretical framework of explainability and its implications on high-stakes biomedical decision-making'. In:* BMC Medical Research Methodology *25.Suppl 1 (2025), p. 282.* DOI: 10.1186/s12874-025-02703-1

Chapter 1 framed the opacity of contemporary AI systems — particularly FMs and LLMs — as an epistemic and societal problem, one that undermines accountability, trust, and the very notion of understanding. It argued that the field of XAI has reached a conceptual impasse: abundant in techniques but lacking a unifying account of what an explanation *is*. This chapter takes up that challenge by laying the theoretical foundations for moving beyond ad hoc tools toward a principled science of explainability. It reconstructs the core dimensions introduced in the triple frontier — faithfulness, intelligibility, and alignment — into a formal and integrative framework that will underpin all subsequent methodological and empirical developments in this thesis.

## 3.1 Reconnecting the Dots of Explainability

Despite significant advances in developing models that are explainable by design [23, 205, 75] and techniques that retrofit explainability onto opaque architectures [147, 112, 10], the field of XAI remains conceptually fragmented. A unifying theoretical substrate—capable

of grounding diverse approaches in shared principles—is still missing. Foundational questions about what constitutes an explanation, how it should be evaluated, and the role of human users in its formation and validation remain under-theorized. In the absence of such foundations, the literature proliferates into parallel efforts, each addressing only isolated aspects of the explainability problem. A prominent symptom of this fragmentation is terminological inconsistency. The terms "interpretable" and "explainable" are often used interchangeably or contrastively, without reference to a common conceptual core [62, 30, 127]. This lexical ambiguity impedes cumulative progress: even when novel methods are proposed, they rarely connect to a shared theoretical framework that would allow systematic comparison, synthesis, or empirical validation across studies. Crucially, there remains no general formalism describing what an explanation for an automated system *is*, nor how the process of constructing and interpreting explanations should be represented. The lack of such a formal vocabulary constrains both the full understanding of individual studies and the interoperability of their insights.

To address such a gap, this work introduces a novel theoretical framework that delineates the core components of the "explainability machinery." Its purpose is not to impose a doctrine but to offer a conceptual reference point that organizes current methods and guides future design. At the center of this proposal lies a new definition of *explanation*. Drawing on philosophical and sociological traditions, an explanation is defined as the interaction between two complementary components: *evidence* and its *interpretation*. The *evidence* comprises information derived from a model—such as parameters, gradients, or rules—while the *interpretation* denotes the semantic meaning that a human assigns to that evidence. For instance, in attention mechanisms, attention values serve as evidence, while the fact that they point to high-relevance features in the input constitutes an interpretation of this evidence. An explanation thus emerges through the act of mapping factual artifacts onto human-understandable meaning, transforming computational traces into insight.

This conception naturally highlights two essential properties of explanations: *faithfulness* and *plausibility*. Faithfulness refers to the degree to which the interpretation accurately reflects the model's causal reasoning [82]. Plausibility, by contrast, measures how well an explanation aligns with user intuition or expectation [82, 81]. Both are relevant, yet they differ in epistemic status: plausibility supports user trust, whereas faithfulness anchors validity. Without faithfulness, plausible explanations risk degenerating into persuasive but misleading narratives—*fictions of transparency* that obscure rather than illuminate model behavior. The framework, therefore, takes faithfulness as the non-negotiable foundation of explainability, with plausibility as a pragmatic complement.

This distinction enables systematic analysis of explanation methods across architectures. For example, attention mechanisms [10, 181], Gradient Class Activation Map (Grad-CAM) [165], and SHAP [112] produce different evidential artifacts—weights, gradients, and Shapley values—whose interpretive mappings can be evaluated for faithfulness within a unified schema. Crucially, the framework also applies to models often presumed "intrinsically interpretable," such as linear regressions or fuzzy logic systems. Even in these cases, explainability depends not only on mathematical simplicity but also on how the quantitative structure is rendered semantically meaningful to users. In short, no model is transparent by default; explainability arises only when evidence is successfully translated into human sense-making. By reconnecting these strands, the proposed framework offers a principled foundation for analyzing, comparing, and ultimately designing explanations that are both faithful to the model and meaningful to the user.

## 3.2 Rethinking Explainability from the Ground Up

The framework sketched in the previous section established a formal account of explanation as the interaction between *evidence* and *interpretation*, anchored in the principles of *faithfulness* and *plausibility*. Yet formal definition alone does not resolve the deeper conceptual tension underlying XAI: what does it mean for a system to be *explainable* in the first place? The field still operates within two narrow paradigms. The first, *structural*, assumes that explainability is an intrinsic property of specific model classes—transparent by design. The second, *procedural*, treats explainability as a post-hoc supplement added to opaque systems. Both approaches reduce explainability to a technical feature rather than an epistemic function, neglecting the conditions under which computational reasoning becomes meaningful to humans.

**This section therefore rethinks explainability as a design-based epistemic practice.** If an explanation arises from the transformation of evidence into interpretation, then explainability concerns the broader system of design choices that make such transformation possible and reliable. It is not merely a property of models or a matter of communication, but a **constitutive principle** that organizes the relation between algorithmic reasoning and human understanding. To rethink explainability *from the ground up* is thus to examine the conceptual grounds—cognitive, epistemic, and socio-technical—on which intelligibility is built. This reconceptualization serves three purposes. First, it situates explainability as **relational**: it depends on the alignment between machine representations and human interpretive capacities. Second, it treats explainability as **continuous**: not a binary attribute but a gradient arising from the interplay among model design, the explanatory interface, and user cognition. Third, it elevates explainability to a **design imperative**: a guiding principle for constructing systems that are intelligible, trustworthy, and ethically aligned. Together, these principles transform explainability from a secondary concern into a primary organizing logic for AI development.

### 3.2.1 Explainability, Explanation, and Interpretation

If explainability is conceived as a design-based epistemic practice, then **interpretation** becomes its central mechanism. Following the core of the framework introduced earlier, an explanation emerges through the translation of *evidence* into *meaning*. Yet to sustain this view, the notion of interpretation itself must be refined. Etymologically, to "interpret" means both "to explain" and "to assign meaning" [35]. The second sense—assigning meaning to evidence—captures the epistemic act that transforms computational traces into intelligible insight. Interpretation is therefore not a supplement to explanation but its constitutive process: the locus where algorithmic reasoning and human understanding converge.

Consider a regression model in which the coefficient on the feature "age" is large and positive. The numerical value alone does not constitute interpretation; interpretation arises when a user understands this as "age substantially influences the outcome." What is being explained, then, is not the number but the relation between model evidence and human meaning. This view rejects the traditional binary between "interpretable" and "post-hoc explainable" models. Both rely on the same epistemic process—bridging the factual and the meaningful—through interpretive design. Explainability, under this account, is inherently **relational**: it exists only through the interaction between computational systems and human agents. Designing for explainability thus entails designing for interpretation.

### 3.2.2 Explainability as a Continuum

If interpretation anchors explainability in relational meaning-making, it follows that explainability cannot be understood as a fixed property that some models possess and others lack. Instead, it should be treated as a **continuum of intelligibility**. Molnar [124] distinguishes between *intrinsic interpretability* and *post-hoc explainability*, but this dichotomy obscures the fact that intelligibility is always co-produced by model design, explanation method, and user cognition. The same system may appear transparent to an expert and opaque to a novice; the same visualization may clarify or confuse, depending on context and expectation.

In this light, **every model possesses some degree of explainability**, determined by how its representational structure, explanatory interface, and user understanding interact. Explainability is not an intrinsic property, but rather a dynamic relationship distributed across the system and the observer. Complex architectures—such as deep networks—can be rendered intelligible through careful visualization or counterfactual analysis, while even simple models can become opaque when stripped of communicative context. The challenge, then, is not to identify where explainability resides, but to *design* the conditions under which it scales with both computational and cognitive complexity. Understanding explainability as a continuum foregrounds its procedural nature: intelligibility is achieved through the ongoing calibration between system transparency and human sense-making.

### 3.2.3 Explainability as a Design Imperative

If explainability is relational and continuous, it must also be **intentional**. It cannot be appended as an afterthought to technical systems but must instead be embedded as a **foundational design principle**. From this perspective, explainability constitutes what software engineering terms a *non-functional requirement* [22]: a condition that shapes architecture, interaction, and evaluation across the entire AI lifecycle. Treating it as such shifts the emphasis from describing models to designing systems that are *intelligible by construction*.

This orientation exposes opacity not as an inevitable consequence of model complexity but often as a failure of design. The well-known example from Ribeiro et al. [147]—where a classifier learned to associate "snowy background" with "dog"—demonstrates that, in the absence of explicit explanation mechanisms, spurious correlations remain invisible until they erode validity and trust. Designing for explainability entails building safeguards against such epistemic blind spots: iterative evaluation, participatory feedback, and reflexive alignment between system reasoning and human understanding.

An explainable system, therefore, is not merely one that *can* be explained but one that *has been designed to support explanation as a mode of interaction and inquiry*. Explainability becomes an organizing logic for system development—linking reliability, ethical accountability, and user intelligibility into a single design imperative. This framing completes the shift from explainability as a descriptive attribute to explainability as an epistemic architecture.

### 3.2.4 Explaining the Data (and the World)

If model explainability concerns how systems reason, data explainability concerns what they reason about. Embedding explainability in system design addresses only half of the challenge. Equally crucial is clarifying what exactly is being explained: the *model*, the *data*, or the *world* that the data represent. Most XAI research has been model-centric,

focusing on elucidating the algorithm's internal logic—its parameters, attention maps, or decision pathways—to ensure **faithfulness**. Yet this focus is incomplete. A model's reasoning cannot be fully understood without understanding the nature of the data on which it reasons. Explaining a model without explaining its data is akin to analyzing syntax without semantics. Explaining data means reconstructing the epistemic, methodological, and social processes through which data is produced. Datasets are not neutral inputs but **constructed representations** of the world, shaped by specific measurement choices, institutional contexts, and value judgments. They embed assumptions about what is worth observing, whose experiences are included, and which phenomena remain invisible. To explain the data is to make these assumptions explicit—to reveal how raw phenomena are transformed into quantitative form and how such transformations delimit what the model can learn or predict.

This perspective extends the notion of **intelligibility** in XAI beyond model behavior to the evidential grounds of model knowledge. A genuinely intelligible system must render both its decision-making process and its evidential foundation transparent. This requires addressing questions such as: What epistemic assumptions underlie the dataset? What socio-technical processes shaped its composition? What forms of uncertainty or exclusion does it encode? Data explainability thus situates models within their epistemic and ethical environments, enabling users to assess not only whether an explanation is faithful to the model, but whether the model itself is faithful to the world. The implications for **alignment** are immediate. Models trained on biased data may produce explanations that are faithful yet unjust, faithfully reproducing inequities embedded in their training sources [12]. Thus, explainability must operate on two levels: internal (model reasoning) and external (data reasoning). Only by integrating both can we achieve explanations that are not merely faithful, but also justifiable and contextually grounded.

Synthesizing model and data explainability responds to the conceptual fragmentation identified earlier. The first concerns understanding the model's causal structure (**faithfulness**); the second, situating that structure within its socio-epistemic context (**intelligibility** and **alignment**). Achieving this synthesis requires a multidisciplinary effort that draws on philosophy of science, data ethics, and domain expertise. Recognizing this duality prepares the ground for a formal account of inference itself—the transformation of data into conclusions. The next section develops such a formalization, providing both the vocabulary to describe reasoning processes and the conceptual bridge linking faithfulness, intelligibility, and alignment within a single explanatory framework.

## 3.3 A Formal Model of Inference

This section formally characterizes the inference process of a general ML model without imposing any task-specific constraints. Such a characterization will introduce the terminology that substantiates the main components of the proposed framework of explainability, whose details are provided in Section 3.4. To this end, this work defines a ML model $M$ as an arbitrarily complex function mapping a *model input* to a *model conclusion* through a sequential composition of *transformation steps*. The whole characterization is exemplified in Figure 3.1.

**Definition 3.1** (Model Input)**.** It is the set of features for a data point in the dataset, either derived from observations or generated synthetically.

It is worth noting that the set of features may consist of any vector representation of a data point, *e.g.*, pixel colors in an image and sub-word embeddings in a document representation. Thus, this definition is independent of the task to be solved or the data type.

**Definition 3.2** (Model Conclusion)**.** It is the model's final output, representing the result of the last link in the chain of transformations applied to the input.

Sometimes the *model conclusion* is called *prediction*, or *inference*, or *forecast*. Since the aforementioned terms are somehow linked to the task to be solved, *e.g.*, the term "forecast" should be used only for model conclusions that affect the future, I adopt the broader term *model conclusion*. The model's conclusion can thus be anything, depending on the task to be solved, *e.g.*, a class probability (in a classification problem), a word vector (in a next-word prediction task), or a scalar value (in a regression problem).

**Definition 3.3** (Transformation Steps)**.** Overall, the decision-making process of $M$ can be represented as a chain of $N > 0$ causally related transformations of the original model input. This causal chain is enforced by the model design (*e.g.*, the sequence of layers in a neural network's architecture or the flow of a decision tree). I call each stage of this causal chain a "transformation step" and denote it with $s_i$, for $i \in [1, N]$. The transformation steps map the model input to the model output via *transformation functions*.

**Definition 3.4** (Transformation Functions)**.** Each transformation step $s_i$ relates to a set of $n_i$ "transformation functions" $f_{i,m_i}$, where $m_i \in [1, n_i]$ indicates one of the possible learnable functions at $s_i$. Note that, in general, the number of such functions would be infinite, but I discretize it, assuming I am working on a real scenario using some computational machine. The transformation functions are mappings from a feature set $x_{i-1,j}$ to a feature set $x_{i,z}$, with $j \in [1, k_{i-1}]$, $z \in [1, k_i]$ (*i.e.*, the arrows enclosed in the ellipses in Figure 3.1). The number $k_i$ denotes the cardinality of the set of all possible feature sets generated by all possible learnable transformation functions at step $s_i$. These transformation functions are generally opaque to the user in the context of the so-called black-box models. At every step in the chain of transformation steps, the model learns one of the possible transformation functions (*i.e.*, the optimal function according to some learning scheme, highlighted with a solid line in Figure 3.1). That is, the model learns the function $\hat{f}_{i,m_i}$ such that $\hat{f} = \hat{f}_{N,m_N} \circ \ldots \circ \hat{f}_{i,m_i} \circ \ldots \circ \hat{f}_{1,m_1}$ is the overall approximation of the true mapping from the model input to the model conclusion. According to the notation above, I denote the model input as $x_{0,0}$ (or simply $x$) and the model conclusion as $\hat{y}_{N,j}$, with $j \in [1, k_N]$ (or simply $\hat{y}$).

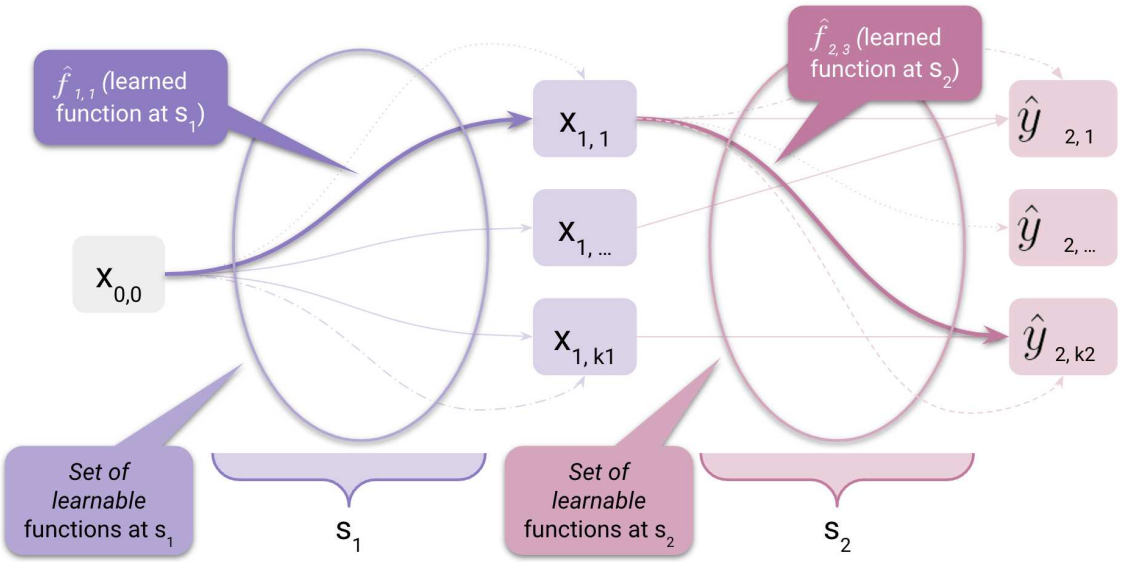



Figure 3.1: Example of transformation functions for two steps $s_i$.

### 3.3.1 Observations

I asserted that, at each transformation step $s_i$, the model picks one function $\hat{f}_{i,m_i}$ among $n_i$ such that $\hat{f}_{i,m_i}(x_{i-1,j}) = x_{i,z}$. This raises issues that increase model opacity. At step $s_i$, the chosen function $\hat{f}_{i,m_i}$ can map different intermediate transformations $x_{i-1,j}$ of the feature set at the previous transformation step into the same transformation $x_{i,z}$ one step further in the chain. This means that the same outcome in the transformation chain, whether intermediate or conclusive, can be achieved through different rationales, making it difficult for a human user to understand which one the model has learned. This can result from a high cardinality of the set of transformation functions and a high complexity of the transformed feature set.

For example, pictures of zebras and salmon can be discriminated against based on their anatomy (*i.e.*, zebras have stripes while salmon have gills) or the environment/habitat (*i.e.*, zebras live in savannas and salmon in rivers). Consider a relatively complex model such as a Convolutional Neural Network (CNN), where a transformation step coincides with a layer within the network architecture. It is generally difficult to understand which kind of transformation $f_{i,m_i}$ this represents, much less if that is human-understandable. Thus, how do I understand which of the $n_i$ possible alternative mappings of $x_{i-1,j}$ led to $x_{i,z}$? This remains an open question with major implications for the discussion around faithfulness, which I will elaborate on in the next section.

## 3.4 Defining Explanations

Recent work on ML explainability produced multiple definitions for the term "explanation". According to Lipton, "explanation refers to numerous ways of exchanging information about a phenomenon, in this case, the functionality of a model or the rationale and criteria for a decision, to different stakeholders" [110]. Similarly, for Guidotti et al., "an explanation is an *interface* between humans and a decision-maker that is at the same time both an accurate proxy of the decision-maker and comprehensible to humans" [65]. Mur-

doch et al. add to how the explanation is delivered to the user, stating that "an explanation is some relevant knowledge extracted from a machine-learning model concerning relationships either contained in data or learned by the model. [...] They can be produced in visualizations, natural language, or mathematical equations, depending on the context and audience" [127]. On a more general note, Mueller et al. state that "the property of being an explanation is not a property of the text, statements, narratives, diagrams, or other material forms. It is an interaction of (i) the offered explanation, (ii) the learner's knowledge and beliefs, (iii) the context or situation and its immediate demands, and (iv) the learner's goals or purposes in that context" [125]. Finally, Miller tackles the challenge of defining explanations from a sociological perspective. The author highlights a wide range of explanations but focuses on those that answer a "why-question" [123].

The definitions mentioned above offer a well-rounded perspective on what constitutes an explanation. However, they fall short in highlighting its atomic components and characterizing their relationships. I synthesize the proposed explanation definition by combining complementary aspects of existing definitions. The result is a concise definition that is easy to operationalize for supporting the analysis of multiple approaches to explainability. The full proposed framework is reported in the scheme in Figure 3.2. The next subsection will formally define its core components, *i.e.*, evidence, interpretations, and explanations. The properties of these components, *i.e.*, explanatory potential, faithfulness, human intuition alignment, and plausibility (with its sub-properties), are examined in §3.5.

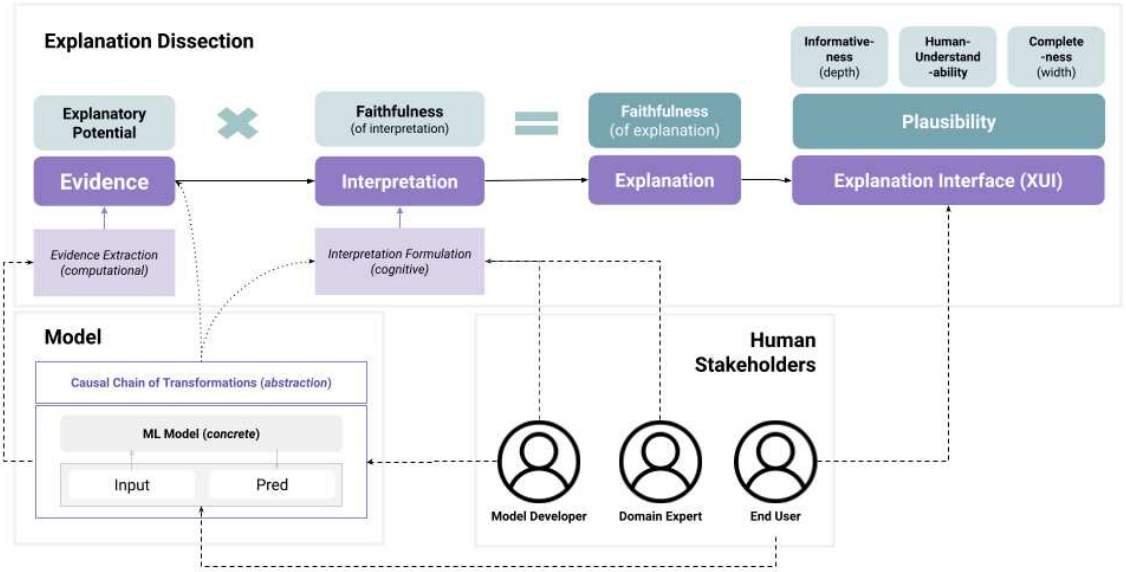



Figure 3.2: Overview of the theoretical framework of explainability.

### 3.4.1 Definitions of the components of the framework

**Definition 3.5** (Explanation)**.** Given a model $M$ which takes an input $x$ and returns a prediction $\hat{y}$, I define "explanation" as the output of an *interpretation function* applied to some *evidence*, providing the answer to a "why question" posed by the user.

**Definition 3.6** (Evidence ($e$))**.** I define "evidence" (denoted by $e$) as whatever kind of objective information stemming from the model I wish to explain, and that can reveal

insights into its inner workings and rationale for prediction (*e.g.*, attention weights, model parameters, gradients, etc.).

**Definition 3.7** (Evidence Extractor ($\xi$))**.** An "evidence extractor" (denoted by $\xi$) is a computational method fetching relevant information about $M$, $x$, $\hat{y}$, or a combination of the three. Then: $e = \xi(x, \hat{y}, M)$. Examples of evidence extractors include *e.g.*, encoder plus attention layers, gradient back-propagation, and random tree approximation. The corresponding extracted evidence includes attention weights, gradient values, and a random tree that mimics the original model. In the peculiar case of a white-box approach, that is, ML models designed to be "easily explainable" by the user (*e.g.*, linear regression, fuzzy rule-based systems), the extraction of evidence is straightforward since all components of the model directly present a piece of semantic information in a human-comprehensible format.

**Definition 3.8** (Explanatory Potential ($\epsilon(e)$))**.** I define "explanatory potential" (denoted by $\epsilon(e)$) of some evidence as the extent to which the evidence influences the causal chain of transformation steps of a model. Intuitively, the explanatory potential indicates "how much" of a model the selected type of evidence can explain. It can be computed either by counting the number of transformation steps affected by the evidence (i.e., *breadth*) or by quantifying the extent to which each individual transformation step is affected (i.e., *depth*).

**Definition 3.9** (Interpretation)**.** An "interpretation" is a function $g$ associating semantic meaning to some evidence and mapping its instances into explanations for a given prediction or the whole model. Then an explanation can be defined as either $E = g(e, x, \hat{y}, M)$, or $E = g(e, M)$, respectively.

**Definition 3.10** (eXplanation User Interface (XUI))**.** I define XUI as the format in which explanations are presented to the end user. This could, for example, take the form of text, plots, infographics, or other visual elements.

### 3.4.2 Observations

**Local vs. Global Interpretations.** Following the existing literature, I relate "evidence" and "interpretation" to the concepts of *locality* and *globality*. Both evidence and interpretations can be either local or global in nature. Local evidence (*e.g.*, attention weights, gradient, etc.) relates relevant model information to a particular model input $x$ and corresponding prediction $\hat{y}$. Global evidence (*e.g.*, full model parameters) is generally independent of specific inputs. It might explain the model's higher-level functioning (providing deeper or broader information) or some of its subcomponents. Similarly, interpretations can provide either local or global semantics for the evidence. A local interpretation of attention could be, *e.g.*, "attention weights are descriptive of input components' importance to model output". On the other hand, a global interpretation of the same evidence may aggregate the heatmaps of all attention weights across the dataset, thereby highlighting specific patterns. For example, in a dog *vs* cat classification problem, a global interpretation of attention may be represented by clusters of similar parts of the animal's body (*e.g.*, groups of ears, tails, etc.) highlighted by the attention activations.

**Generating Interpretations.** Given some evidence involved in one or more steps $s_i$ of $M$, I "guess" how this evidence is involved in the opaque input-to-output transformations

by formulating an interpretation $g$ of some extent of the model's decision-making process. At a low level, I generate a candidate hypothetical function $g$ that encapsulates the approximations $f^*_{i,m_i} \tilde{=} \hat{f}_{i,m_i}$ of the behavior of certain functions learned by $M$ at some steps $s_i$. On an abstract level, interpretations can be seen as hypotheses about the role of evidence in the explanation-generation process. Like a good experimental hypothesis, a good interpretation satisfies two core properties: (i) it is testable, and (ii) it clearly defines dependent and independent variables. Interpretations can be formulated using different forms of reasoning (*e.g.,* deductive, inductive, abductive, etc.). In particular, the survey on explanations and social sciences by Miller reports that people usually make assumptions (*i.e.*, in this work's context, choose an interpretation) via social attribution of intent (to the evidence) [123]. Social attribution concerns how people attribute or explain others' behavior, rather than the actual causes of that behavior. Social attribution is generally expressed through folk psychology, which involves attributing intentional behavior using everyday terms such as beliefs, desires, intentions, emotions, and personality traits. Such concepts may not be the true cause of the described behavior, but they are the ones humans use to model and predict each other's behavior. This may lead to a misalignment between a hypothesized interpretation of some evidence and its actual role in the model's inference process. In other words, reasoning on evidence through folk psychology might generate interpretations that are *plausible* but not necessarily *faithful* to the inference process of the model (such terms will be further explored in Section 3.5).

### 3.4.3 Framework Summary

As depicted in Figure 3.2, the proposed framework operates as follows: an *explanation* is derived from the *interpretation* of certain *evidence*. This evidence is produced by an *evidence extractor* that operates on the model in combination with its inputs and outputs. The expressiveness of the evidence can be characterized in terms of its *explanatory potential* (breadth and depth). This potential, along with the interpretation's adherence to the model's actual inference process (*faithfulness*), directly affects the explanation's overall faithfulness. Finally, the *XUI* is responsible for conveying the explanation to the user.

## 3.5 Concerning Faithfulness and Plausibility

Having defined the atomic components of the framework—evidence, interpretation, and explanation itself—it becomes necessary to examine the qualitative properties that determine their epistemic value. Not all explanations are created equal: some accurately reflect the reasoning of a model, while others merely appear convincing to human observers. The distinction between *faithfulness* and *plausibility* captures this essential divide. Faithfulness concerns the extent to which an explanation corresponds to the model's true causal mechanisms, whereas plausibility reflects how well the explanation aligns with human intuition, expectations, or beliefs. Both are necessary for trustworthy and usable explainability, yet they operate on distinct epistemic planes—truth versus comprehension. This section analyzes how these properties interact, the risks that arise when one is prioritized over the other, and the methodological foundations needed to balance them. By disentangling faithfulness from plausibility, the framework establishes the normative ground for designing explanations that are both reliable representations of model reasoning and accessible to human understanding.

### 3.5.1 Faithfulness of interpretations and explanations

In the previous sections, I observed that social attribution is a double-edged sword for the interpretation generation process, as it may enhance plausibility without ensuring faithfulness. This issue was highlighted by Jacovi & Goldberg, who introduced the property of explanations known as *aligned faithfulness* [81]. In the authors' words, an explanation satisfies this property if "it is faithful and aligned to the social attribution of the intent behind the causal chain of decision-making processes." The proposed framework enables us to advance our characterization of this property. I note that the property of aligned faithfulness pertains only to interpretations, not evidence. The latter has no inherent meaning. Its semantics are defined by an interpretation that may or may not involve attributing social intent to the causal chain of inference processes.

**Definition 3.11** (Faithfulness (of interpretation))**.** Given an interpretation function $g$, describing some transformation steps $s_i$ within a model $M$'s inference process, I want to be able to prove that $g$ is faithful (at least to some extent) to the actual transformations made by $M$ to an input $x$ to get a prediction $\hat{y}$. Namely, I define the property of *faithfulness of an interpretation* $\phi_i(g, e)$ as "the extent to which an interpretation $g$ accurately describes the behavior of some transformation functions $f_{i,m_i}$ that some model learned to map an output $x_{i-1,j}$ at $s_{i-1}$ into $x_{i,z}$ at $s_i$ making use of some instance evidence $e$".

**Definition 3.12** (Faithfulness (of explanation))**.** Given some evidence $e$ and its interpretation function $g$, I say that a related explanation is faithful to some transformation steps if the following conditions hold: (i) the evidence $e$ has explanatory potential $\epsilon_i > 0$, and (ii) the interpretation $g$ has faithfulness $\phi_i > 0$. Then I can define the *faithfulness of an explanation* ($\Phi$) as a function of the faithfulness of each step's interpretation and the explanatory potential of that step.

For example, I could define $\Phi = \sum_i \epsilon_i \phi_i \ \forall i \in I \subseteq [1, N]$ where $I$ is the set of indices of transformation steps $s_i$ that involved the evidence $e$. Thus, the faithfulness of an explanation is the sum of the faithfulness scores of its components, *i.e.*, the faithfulness of the interpretations of the evidence involved in generating the explanation. Additionally, the related explanatory power weights the faithfulness of each interpretation, following the intuition that evidence with a higher $\epsilon$ should have a greater impact on the interpretation's overall faithfulness score. I can have various measures of faithfulness associated with different explanation types, in the same way that I have different metrics to evaluate an ML model's ability to complete a task. Thus, $\phi_i$ is implicitly bounded.

When designing a faithful explanatory method, I can opt for two approaches. I can achieve faithfulness "structurally" by enforcing this property on pre-selected interpretations in model design (*e.g.*, imposing constraints on transformation steps that limit the range of learnable functions). This direction has been recently explored by Jain & Wallace [84] and Jacovi & Goldberg [81]. An alternative, naive strategy is trial and error: formulating interpretations and assessing their faithfulness via formal proofs or requirements-based testing using proxy tasks. While formal proofs are still missing in the current literature, several tests for faithfulness have been recently proposed [4, 83, 193, 166, 115, 37].

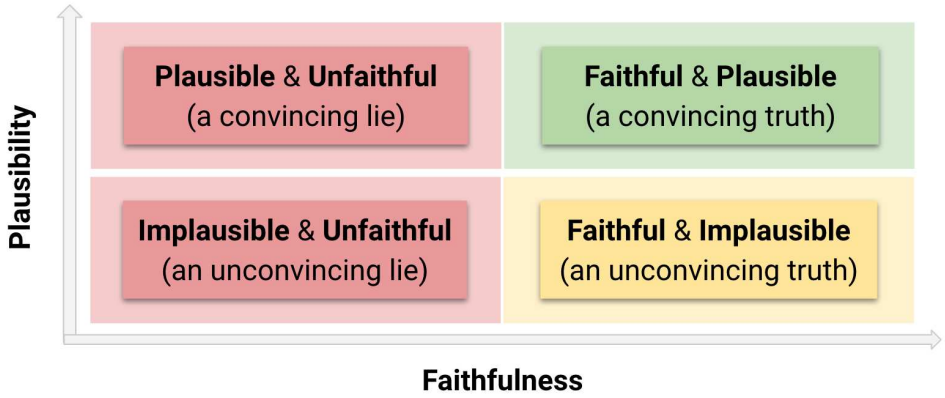



Figure 3.3: Overview of the outcome on the user of the interaction between faithfulness and plausibility.

### 3.5.2 Plausibility of the explanation user interface

Explanations are intended for delivery to specific target users, and that is when XUIs come into play. I argue that an XUI is characterized by three main properties: (i) human understandability, (ii) informativeness, and (iii) completeness. The *human-understandability* is the degree to which users can understand the answer to their "why" question via the XUI. This property depends on user cognition, bias, expertise, and goals, among other factors, and is influenced by the complexity of the selected interpretation function. The *informativeness* (*i.e.*, depth) of an explanation is a measure of the effectiveness of an XUI in answering the why question posed by the user. That is the depth of information for some $s_i$ of great interest in the XUI. The *completeness* (*i.e.*, width) of an explanation is the extent to which an XUI describes the overall model's workings and the degree to which it allows for anticipating predictions. That is the width in terms of the number of $s_i$ the XAI spans. Note that informativeness and completeness are bound by the explanatory potential of the evidence (*e.g.*, attention weights do not explain the entire model, just some transformation steps; in contrast, the complete set of model parameters does). The combined value of the three properties mentioned above of XUIs drives the plausibility of an explanation.

**Definition 3.13** (Plausibility (of explanation))**.** I define *plausibility* as the degree to which an explanation is aligned with the user's understanding of the model's partial or overall inner workings.

Plausibility is a user-dependent property, and as such, it is subject to the user's knowledge, bias, etc. Unlike faithfulness, the plausibility of explanations can be assessed via user studies. Note that a plausible explanation is not necessarily faithful, just like a faithful explanation is not necessarily plausible. It is desirable for both properties to be satisfied in the design of some explanation. Interestingly, an unfaithful yet plausible explanation may lead a user to believe that a model behaves according to a rationale when, in fact, it does not. This raises ethical concerns that poorly designed explanations could spread inaccurate or false knowledge among end users. Figure 3.3 provides a simplified problem overview.

### 3.5.3 Connecting the Dots: Abductive Reasoning

Abductive reasoning, a pivotal concept in XAI, is crucial in generating and interpreting explanations for ML models [72]. This form of reasoning allows one to infer a condition that explains an observed consequence. It is deeply entrenched in the philosophy of science and particularly resonates with the works of Charles Sanders Peirce. Peirce's perspective on abduction as a logical inference that generates new hypotheses is foundational in this context [138]. The application of abductive reasoning spans various domains. In medical diagnostic AI, for instance, it helps infer potential health conditions from symptomatic data and patient history, thereby enhancing the trustworthiness and reliability of AI recommendations in healthcare [100]. In AI, abductive reasoning primarily elucidates the 'why' behind AI decisions and predictions. It involves constructing *plausible* hypotheses from observed data, followed by iterative refinement through testing and validation [118]. Alternatively, from my perspective, it involves constructing interpretations of the evidence that align with human intuition. Explainable methods usually fall short at this point, when human intuition surpasses the importance of faithfulness in the interpretation. I advocate focusing on faithfulness, ensuring that the interpretation accurately reflects the model's decision-making process. *Alignment* with human intuition is a nice-to-have for the interpretation but a necessity for the final explanation, which is why it must be considered when designing the XUI. A significant challenge in XAI is making the abductive reasoning processes *understandable* to users. Designing XAI interfaces that effectively communicate the AI's hypotheses and reasoning enhances user understanding and trust, particularly in high-stakes domains such as healthcare [28].

Finally, abductive reasoning in AI is dynamic, evolving with new data and insights. This necessitates AI systems that can adapt and refine their explanations over time, underlining the importance of continuous learning and adaptation in AI technology. Abductive reasoning's role in ML models is multi-dimensional and vital, extending beyond hypothesis generation to providing clear, understandable explanations, thereby bolstering the usability and reliability of AI systems.

## 3.6 Case Studies: Framing Explainability Strategies

In this section, I apply the theoretical framework to existing *explainability* methods. Additionally, I demonstrate the application of this framework to two "easily explainable" methods: linear regression and fuzzy models. The objective is to provide concrete examples of how various XAI methods fit well within the proposed framework's components. I also want to emphasize the usefulness of a common set of terms to indicate different concepts frequently encountered in XAI discourse. Despite the examples being limited to some of the most influential methods, I believe they will provide sufficient guidance for future research that applies this framework.

### 3.6.1 Attention

The introduction of attention mechanisms has been one of the most notable breakthroughs in DL research in recent years. Originally proposed to improve neural machine translation [10], attention mechanisms now underpin many SotA architectures across a wide range of tasks, most notably as the core component of the Transformer model [181].

At a high level, the simplest self-attention mechanism can be described as a three-step causal process: (i) *encoding*, (ii) *weighting* of the encoded representations via attention scores, and (iii) *decoding* into the final model output.

During the encoding phase, the input features are represented as a sequence of $t$ tokens, which are then projected into three distinct vector spaces: the *queries* ($Q$), the *keys* ($K$), and the *values* ($V$). In the self-attention setting, these three matrices are derived from the same input sequence, i.e., $Q = XW_Q$, $K = XW_K$, and $V = XW_V$, where $W_Q$, $W_K$, and $W_V$ are learned projection matrices.

The attention mechanism computes a set of similarity scores between each query and all keys, typically using a scaled dot-product:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V.$$

The resulting normalized scores represent the degree to which each token should pay attention to every other token in the sequence. These scores—commonly referred to as *attention weights*—serve as a measure of relational importance among tokens. The final step combines the weighted representations to produce contextually enriched embeddings, which are then decoded or passed to downstream layers.

Formally, the function learned by the model can be described as a composition of transformation steps:

$$\hat{f} = f_{3,m_3} \circ f_{2,m_2} \circ f_{1,m_1},$$

where each $f_i$ corresponds to one of the causal transformations in the chain.

**Evidence.** For an input $x$ split into $t$ tokens, let $f_{1,m_1}$ denote the encoding step such that $f_{1,m_1}(x) = \bar{X}$, the set of embedded token representations. The subsequent transformation $f_{2,m_2}$ computes a weighted combination of these embeddings:

$$f_{2,m_2}(\bar{X}) = \sum_{j=1}^{t} \alpha_j \, \bar{x}_j,$$

where $\alpha_j$ is the attention weight assigned to token $\bar{x}_j$. The set of these weights constitutes the *evidence*:

$$e_{\text{att}} = \{\alpha_j \mid f_{2,m_2}(\bar{X}) = \sum_{j=1}^{t} \alpha_j \, \bar{x}_j\}.$$

The explanatory potential $\epsilon(e_{\text{att}})$ can be estimated as the ratio of the number of parameters in the analyzed attention layer to the total number of model parameters, capturing how much of the model's reasoning this evidence reveals.

**Interpretation.** The interpretation of this evidence is a function $g_{\text{att}}(e(x, \hat{y}))$ that describes how the attention weights contribute to the model's output—namely, how the weighted embeddings are decoded into the model's final conclusion. A widely adopted interpretation is that tokens receiving higher attention weights are more influential for the model's prediction.

**Faithfulness.** While the above interpretation is intuitively appealing and thus *plausible*, its *faithfulness* has been extensively debated. Several studies have shown that attention weights do not necessarily correspond to causal importance or contribution to model decisions [83, 166, 194]. In other words, the distribution of attention scores can often be modified without significantly affecting the model's output, indicating that attention—though

informative—does not inherently provide a faithful explanation of model reasoning. Consequently, the explanatory role of attention mechanisms remains an open and contested question within the broader landscape of XAI.

### 3.6.2 Grad-CAM

Gradient-weighted Class Activation Mapping (Grad-CAM) [165] is a widely used technique for visually explaining the predictions of convolutional neural networks (CNNs). It leverages the gradients of the target class with respect to the final convolutional layer to generate coarse localization maps that highlight the regions of the input image most relevant to a particular decision. In essence, Grad-CAM produces a *class-discriminative localization map* that indicates which parts of an input image contribute positively or negatively to a given class prediction.

For a given input $x$, let $A^k$ denote the $k$-th feature map of the last convolutional layer, and let $y^c$ be the score for class $c$ (before the softmax). Grad-CAM computes the importance weight $\alpha_k^c$ for each feature map by averaging the gradient of the class score with respect to that feature map:

$$\alpha_k^c = \frac{1}{Z} \sum_{i,j} \frac{\partial y^c}{\partial A_{ij}^k},$$

where $Z$ is the number of spatial positions in $A^k$. The final localization map for class $c$ is then obtained as:

$$L_{\text{Grad-CAM}}^c = \text{ReLU}\left( \sum_k \alpha_k^c A^k \right),$$

where the ReLU ensures that only features that positively influence the class score are visualized. The resulting heatmap highlights regions in the input image that contribute most strongly to the model's decision.

**Evidence.** The evidence extracted by Grad-CAM, denoted as $e_{\text{grad}} = \xi_{\text{grad}}(M, x)$, consists of the feature activation maps $A^k$ from the final convolutional layer and the corresponding gradient-derived importance weights $\alpha_k^c$. These quantities encode how changes in each spatial feature location influence the prediction for class $c$. The explanatory potential $\epsilon(e_{\text{grad}})$ depends on the proportion of the model's parameters and layers involved in the generation of these activations relative to the total model size—similar to how attention weights capture a localized but partial view of model reasoning.

**Interpretation.** The interpretation function $g_{\text{grad}}(e(x, \hat{y}))$ maps the evidence to a spatial heatmap showing which regions of the input image most strongly affect the output for a given class. Under this interpretation, larger activation values in the heatmap indicate greater relevance or contribution to the final prediction. Thus, Grad-CAM provides an intuitive visualization linking internal model features to semantically meaningful image regions.

**Faithfulness.** The faithfulness of Grad-CAM explanations has been empirically assessed using occlusion-based tests [4]. In these evaluations, parts of the input image are systematically masked, and the resulting change in the model's output is compared with the saliency indicated by the Grad-CAM heatmap. A high correlation between predicted importance and the observed impact on model confidence suggests strong faithfulness. However, while Grad-CAM often produces visually convincing and class-consistent explanations, its resolution is limited by the spatial granularity of the final convolutional layer, and its faithfulness may vary across architectures and layers.

### 3.6.3 SHAP

Lundberg and Lee (2017) introduced SHAP [112], a method for assigning an importance value to each feature used by an opaque model $M$ to explain an individual prediction $\hat{y}$. SHAP generalizes several earlier explanation approaches, including Local Interpretable Model-agnostic Explanations (LIME) [147], DeepLIFT [168], layer-wise relevance propagation [9], and classical Shapley value estimation.

Formally, SHAP approximates the local behavior of a complex model with a simple additive model:

$$g(z') = \phi_0 + \sum_{i=1}^{M} \phi_i z'_i, \tag{3.1}$$

where $z' \in \{0,1\}^M$ represents a simplified binary version of the input $x$ indicating the presence or absence of each feature, $M$ is the number of features considered, and $\phi_i \in \mathbb{R}$ is the Shapley value quantifying the contribution of the $i$-th feature to the model output. The intercept $\phi_0$ corresponds to the expected value of the model's output over the background dataset.

**Evidence.** The evidence used by SHAP is the set of model outputs produced for perturbed versions of the input $x$ within its local neighborhood:

$$e_{\text{shap}} = \xi_{\text{shap}}(M, x) = \{M(x') \mid x' \in \mathcal{N}(x)\},$$

where $\mathcal{N}(x)$ denotes the neighborhood of sampled instances derived from $x$ by masking subsets of its features. These local predictions form the factual basis for computing the additive importance values $\phi_i$. The explanatory potential $\epsilon(e_{\text{shap}})$ can be expressed as the ratio between the number of sampled predictions used to estimate the Shapley values and the total number of possible feature subsets. Intuitively, the larger the explored neighborhood, the greater the explanatory coverage.

**Interpretation.** The interpretation function $g_{\text{shap}}$ maps the extracted evidence $e_{\text{shap}}$ to a locally interpretable additive model $g(\cdot)$ as defined in Equation 3.1. Through this model, the contribution of each feature to the final prediction can be quantified via its corresponding $\phi_i$ value. Thus, analyzing $h(z')$ (or equivalently $g(z')$) provides a local explanation $E_{\text{shap}}$ of how the original model $M$ transforms input features into its output for the given instance $x$.

**Faithfulness.** While SHAP does not directly provide a numerical measure of faithfulness, its theoretical foundation is built on three desirable properties that implicitly enforce it. The first, *local accuracy*, ensures that the additive explanation model exactly matches the output of the original model for the given instance. The second, *missingness*, guarantees that features absent from the simplified input ($z'_i = 0$) have zero contribution, i.e., $\phi_i = 0$. The third, *consistency*, ensures that if a model changes such that a feature's marginal contribution increases (while all others remain fixed), its corresponding $\phi_i$ does not decrease. Lundberg and Lee demonstrated that SHAP is the only additive feature attribution method that satisfies all three properties simultaneously. Together, these axioms define a requirements-based notion of faithfulness that is consistent with the framework described in § 3.5.

### 3.6.4 Linear regression models

Linear regression models are not an explanation method, but they are typically considered *intrinsically interpretable*. Following the proposed framework, I argue that defining them,

along with other models, as *intrinsically interpretable* is inaccurate and often misleading. The definition of what is simple for humans to interpret is not well-defined. I can provide various examples of models that are easy for practitioners to interpret but are almost impenetrable to non-expert users.

A linear regressor $\hat{f}_{lin}(\cdot)$ is typically formulated as:

$$\hat{f}_{lin}(x) = \beta_0 + \sum_{i=1}^{N} \beta_i x'_i \tag{3.2}$$

where $\beta_i$ are the weights of the learned features, $N$ is the feature space dimension, and $x'_i$ denotes normalized $x_i$.

**Evidence.** The implicit assumption, claiming that a linear model is intrinsically interpretable, is that the weights $\beta_i, 1 \leq i \leq N$ are a good explanation for the model. Thus $e_{lin} = \{\beta_i\}_1^N$. We have the maximum explanatory potential $\epsilon(e_{lin})$ with a linear model because we can fully describe the model with $e_{lin}$.

**Interpretation.** Assuming a normalization of the features, we can say that the higher the value of $\beta_i$, the higher the contribution of the feature $x_i$ to the model prediction.

**Faithfulness.** There are no doubts about the faithfulness of the interpretation of the predictions given the normalization assumption, and in fact, a linear model is normally considered an intrinsically interpretable method. However, a real scenario does not guarantee its plausibility to a non-expert user.

### 3.6.5 Fuzzy models

Fuzzy models, especially in the form of Fuzzy Rule-Based Systems (FRBSs), represent effective tools for modeling complex systems using a human-comprehensible linguistic approach. Owing to these characteristics, they are generally regarded as white or gray boxes and are often considered good options for interpretable AI [54]. Although a detailed description of fuzzy modeling goes beyond the scope of this thesis, it is important to specify that FRBSs perform their inference (*i.e.*, calculate a conclusion) by exploiting a knowledge base composed of linguistic terms and rules. Thanks to this linguistic approach and the fact that fuzzy set theory can naturally embed uncertainty and vague concepts, FRBS is generally considered *intrinsically interpretable* models. A fuzzy rule is usually expressed as a sentence in the form:

```
IF <antecedent> THEN <consequent>
```
(3.3)

where `antecedent` is a logic formula created by concatenating clauses like `` `X IS a' `` with some logical operators, where `T` is a linguistic variable (associated with one input feature) and $a$ is a linguistic term. Thanks to this representation, the antecedent of each rule provides an intuitive and human-understandable characterization of a specific class/group.

The form of `consequent` varies according to the type of model and fuzzy reasoner used. Still, it can be seen as a function that calculates the model's conclusion, such that the more a sample satisfies the antecedent, the higher the rule's weight in the final calculation. Note that, due to the fuzziness of the model, all rules can be applied simultaneously, although with different weights.

**Evidence.** The rules are good evidence for a large part of the model: they characterize the feature space using a self-explanatory formalism that human operators can read and validate. The fuzzy terms are implemented as fuzzy sets with corresponding membership

functions, typically parametric curves such as triangular, trapezoidal, sigmoidal, or Gaussian.

**Interpretation.** The fuzzy sets used to create the fuzzy terms and evaluate the satisfaction of the antecedents have self-explanatory interpretations: they define how much a value belongs to a given set employing membership functions. The fuzzy rules are also self-explanatory. The only part that requires a proper interpretation is the output calculation function. In the case of Sugeno reasoning, such functions can be viewed as linear regression models; hence, all considerations discussed in Section 3.6.4 remain valid in the context of fuzzy models.

**Faithfulness.** Similarly to linear regression models, there is no doubt about the faithfulness of the interpretation of predictions after a normalization step. However, in the case of special transformations (*e.g.*, log-transformation), some of the intrinsic interpretability might be lost in favor of better fitting to training data [54]. Since it is often the case that features in biomedicine (see, *e.g.*, clinical parameters) follow a log-normal distribution, such transformations are very frequent and delicate.

### 3.6.6 Large Language Models

The proliferation of LLMs presents a unique and formidable challenge for explainability. Their scale, emergent capabilities, and the distributed nature of their internal representations make them a quintessential "black box." An LLM's inference process is a deep chain of transformation steps, where each Transformer block refines a set of token representations through self-attention and feed-forward computations. Applying our framework here reveals the acute tension between plausible and faithful explanations.

**Evidence.** The sheer size of LLMs offers a vast landscape of potential evidence ($e$). An evidence extractor ($\xi$) can pull information from various points in the model's computation for a given input $x$. This evidence can manifest in several forms. For instance, one can examine the *attention patterns* across tens or hundreds of attention layers, yielding a high-dimensional tensor of weights that captures how tokens relate to one another from the different "perspectives" of each attention head. Another form of evidence is the *neuron activations* within the feed-forward network layers, which are often investigated for mechanistic interpretability to identify neurons that correlate with specific concepts [44]. The *hidden states*, or the output vectors for each token at the end of each Transformer block, serve as evidence of the model's evolving contextual understanding. Furthermore, *gradients* of the loss with respect to input embeddings provide saliency evidence, indicating which input tokens are influential. Perhaps most uniquely, the LLM itself can be prompted to produce a *model-generated rationale*, a natural language string that purports to be an explanation of its own reasoning. This text constitutes a distinct and highly compelling form of evidence. However, we treat rationales as evidence only in the weak sense of 'model-produced artifacts,' acknowledging they may be unfaithful to the internal computation.

**Interpretation.** Each type of evidence is associated with a common, intuitive interpretation ($g$). For attention patterns, the interpretation remains that weights signify "importance" or "relatedness" between tokens. For neuron activations, the interpretation is that individual neurons or small circuits correspond to concepts that are understandable to humans. The interpretation of gradients or saliency maps is that higher values correspond to greater importance of the input feature. The interpretation of model-generated rationales is straightforward: the text accurately describes the causal reasoning steps the model took

to arrive at its conclusion.

**Faithfulness.** Evaluating the faithfulness of these interpretations is a central open problem in XAI, a task made exceptionally difficult by the scale of LLMs. As discussed in §3.6.1, the interpretation of attention as a faithful explanation of importance has been widely challenged. For neuron activations, faithfulness is often assessed via causal intervention, such as ablating a neuron to determine whether a model's behavior changes as predicted; however, concepts are often represented in distributed form, which complicates this analysis. Most critically, the interpretation of model-generated rationales faces a profound challenge to faithfulness. Research has shown that these rationales are often post-hoc justifications rather than a true reflection of the model's internal process [207]. The model learns to generate plausible-sounding text that correlates with the correct answer, but this text may have no causal link to the actual prediction mechanism. The LLM is a master of plausibility, but this plausibility offers no guarantee of faithfulness.

This case study perfectly illustrates the core argument of this thesis: LLMs exemplify the danger of conflating plausibility with faithfulness. They can generate explanations that are maximally plausible to a human user (in natural language text) while potentially being unfaithful to the model's actual, sub-symbolic inference process. Our framework, by forcing a clear separation between **evidence** (the generated text) and a **faithful interpretation** (does this text reflect reality?), provides the necessary critical lens to navigate the complex and often deceptive landscape of LLM explainability.

## 3.7 The Role of the User

The framework developed in the previous sections has primarily focused on the internal mechanics of explainability—how evidence, interpretation, and explanation interact within a model to produce intelligible outcomes. Yet, explainability is never an entirely technical property; it is an epistemic and communicative process that necessarily involves a human participant. Models do not explain themselves in isolation: explanations are produced, interpreted, and validated through interaction with users. The human thus occupies a constitutive, rather than peripheral, role in the explainability ecosystem. This section, therefore, shifts the analytical focus from the model to the user, examining how human actors engage with, co-produce, and assume responsibility for explanations. It explores questions of *ownership*—who generates and who understands explanatory content—and introduces a relational model of explainability in which the *explainer*, the *explaining*, and the *explainee* form an interconnected triad. By re-centering the user in the explanatory loop, this section highlights that explainability is not merely a property of systems, but a shared human–machine achievement.

### 3.7.1 Evidence, Interpretations, and Explanations: Who Owns What?

The concept of ownership in explanations is multifaceted and central to understanding explainable ML. In the so-called "intrinsically explainable models" (or, as I discussed, "more easily explainable models"), the system naturally generates a component of the explanations as part of its processing. This is the evidence, factual data that emerges directly from the model's decision-making process. The ownership of this objective information belongs to the model, as it is a model's product or by-product. However, identifying tangible data from the model as evidence is part of the human process of designing the explanation. While the model generates the initial evidence, the final understanding and

contextual arrangement of this data are undertaken by human users. This implies dual ownership. The system 'owns' the initial explanatory data. It provides an opportunity for human stakeholders to identify this data as the evidence supporting an explanation. Identifying the evidence is the human (owned) process of relating some intuitive explanatory potential to information stemming from the model.

Similarly, the ultimate interpretation and the meaning derived from this evidence are 'owned' by the human users. In this context, ownership is also closely tied to responsibility. Human users are responsible for interpreting evidence within the context of their domain knowledge and the specific situation at hand. To tap into the philosophy of linguistics and borrow from the basics of the semiotic triangle [174], I might refer to the data itself as the *signifier*, the evidence as the *referent*, and the interpretation as the *signified*. The semiotic triangle illustrates the relationship between a concept (the "signified"), the physical form or symbol that represents it (the "signifier"), and the real-world object or idea to which it refers (the "referent"). This highlights how meaning is constructed in human language and thought. For a real-world example in the context of ML explainability, the semiotic triangle can be applied to the concept of a "feature importance" score in a decision tree model. In this case, the signified is the concept or idea of "feature importance" (an interpretation), which represents the relative importance or contribution of an input feature to the prediction made by the model; the signifier is the numerical score or graphical representation, *e.g.*, a bar chart, (the data) used to indicate the importance of each feature in the model's decision-making process; the referent is the actual data attribute or input variable to which the feature importance score is applied (the evidence), influencing the model's predictions in the real world.

This shared-ownership representation raises critical questions about the design of AI systems and the type of explanations that should be generated. It necessitates a design approach that considers not only the technical capability of AI systems to provide explanatory data but also the ability of human users to recognize and interpret this data effectively.

Ownership becomes even more nuanced when considering models that are harder to explain, the so-called "black-box" models. These models, such as deep neural networks, do not naturally provide easy access to explanations of their decisions. As a result, external methods, known as post-hoc explainability techniques, are employed to interpret the model's behavior. Here, the ownership of explanations is even more distributed. The AI system still 'owns' the decision-making process, but does not inherently own all the explanatory data, as these are not entirely products or byproducts of the model. Instead, explanations are generated using additional tools and methods that derive explanatory information from initial model data and are then interpreted (*e.g.*, SHAP). These tools aim to provide a proxy context for examining and explaining the model's decision-making process, but they introduce an additional layer of abstraction and potential bias. Developers of these tools are responsible for ensuring that the post-hoc methods accurately represent the model's decisions, while users must critically assess the validity and relevance of these explanations. In this case, the ownership of the explanation is shared among the AI model (raw evidence), the explainability tool (processed evidence), and the human interpreters.

### 3.7.2 Explainers, Explainings, and Explainees

In the academic exploration of ML explainability, a progressive understanding of the roles of the *explainer* and the *explainee* has emerged, emphasizing the dynamic, interactive nature of this relationship. The discourse initially focused on the need for *explainers*

(AI systems) to provide *explainees* (human users) with comprehensible, relevant explanations, particularly in high-stakes domains such as healthcare. Early research underscored explainers' responsibility to bridge the gap between complex ML models and the practical needs of domain experts, thereby enhancing the explainee's ability to make informed decisions. The narrative has evolved with the introduction of frameworks that shifted towards interactive explainability. For instance, Rovolis et al. recently explored the impacts of participatory design on data-driven decision-making in organizations, emphasizing the integration of participatory activities to achieve better outcomes [158]. This approach blurred the traditional boundaries between explainer and explainee, suggesting a more collaborative and iterative understanding of ML models. The concept of personalized explainability emerged, underscoring the importance of tailoring explanations to the individual needs and contexts of explainees. This perspective acknowledges the diversity of the user base for automated systems and the necessity for explainers to tailor their outputs accordingly. For example, Gould et al. investigated patients' views on AI for risk prediction in shared decision-making for knee replacement surgery, emphasizing the need for AI tools to provide personalized information that empowers patients and supports a partnership between clinicians and patients [60].

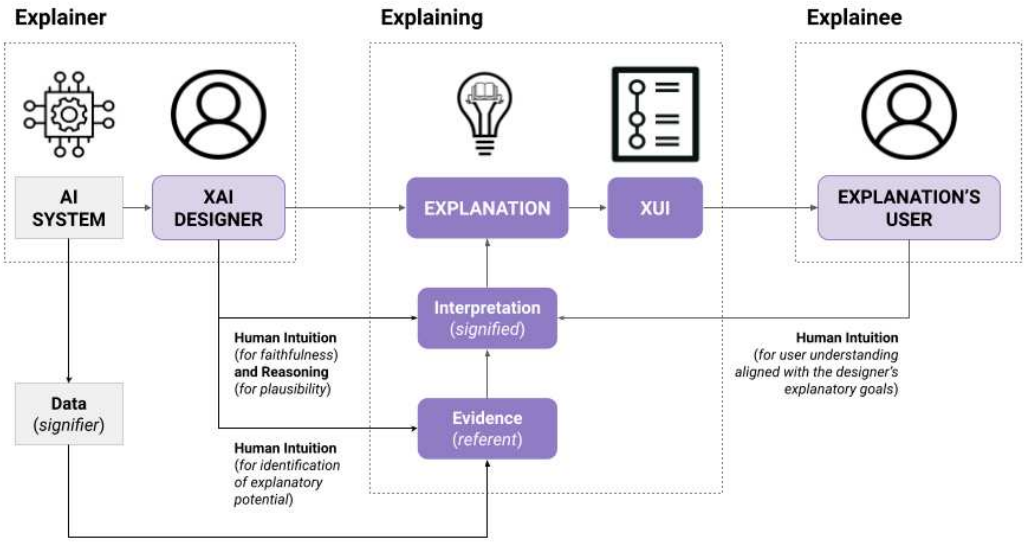



Figure 3.4: The proposed relationships among the explainer, the explaining, and the explained in ML explainability.

Further contributions focused on evaluating explainers in educational contexts and designing user studies to assess the effectiveness of explanations. These studies emphasized the explainer's role in being comprehensible to non-expert users and the active role of the explainee in editing and understanding ML models. The conventional wisdom that ranks ML algorithms by explainability was challenged, advocating a more nuanced, user-centered approach. This shift recognized the varied needs and contexts of explainees, suggesting that the effectiveness of an explainer should be assessed based on its relevance and utility to the specific user. Most recently, the integration of explainable AI methods into the learning loop and the potential of explanations for eliciting user control and feedback were discussed. These studies highlighted the evolving nature of the explainer-explainee

relationship, in which explanations serve as a two-way communication channel, enhancing both model performance and user understanding. In conclusion, the scholarly work in this field collectively advances understanding of the roles of explainer and explainee in ML explainability. They reflect a shift from a unidirectional flow of information to a more interactive, user-centric model, recognizing the importance of context, personalization, and active user engagement in explainable AI.

Based on the contribution of the present work, grounded in the proposed framework, I suggest a shift in perspective. Such a proposal is synthesized in Fig. 3.4. The explainee remains the user of the human explanation. On the other hand, I argue that the explainer is not only created by the AI system but also by an admixture of human knowledge, design, and raw data. Unlike previous work, I focus on the role of the explainer as a human-machine pair in which the human plays an active role. The human is the designer of explainability for a given ML model. This challenge may be more or less difficult depending on the circumstances (*e.g.*, the model type to explain, the degree of designer intervention in the model architecture, etc.), and I will discuss this in-depth in the next section.

Before that, I need to define a novel third component to enrich the discussion around the core roles of the explainability pipeline. I call this component the "*explaining*", which has the concrete role of explaining. The "explaining" embeds the information the explainee needs to understand to grasp the model's decision-making process. To the author's knowledge, this role is usually understated and remains a simple by-product of human design. I argue that this component deserves its own space, as it is crucial to the discussion around explainer-to-explainee knowledge communication. The "explaining" involves the explanation, constructed from its atomic components previously discussed in this thesis (evidence and interpretation), and the final overlay of the explanation user interface that directly relates to the explanation target user. The human role is key to defining "the explaining". First, the XAI designer must use human intuition to identify the explanatory potential in the evidence. This process converts raw data from the AI system into actionable information, enabling the generation of an explanation. Second, intuition and human reasoning[1] are needed to formulate an interpretation. Third, the explanation user must resonate with the interpretation, meaning that he must be able to use human intuition to understand (at a certain, not necessarily full, degree) the rationale for an explanation [25]. This returns to the debate between faithfulness and plausibility that I discussed in §3.5.

The dynamics between the explainer and the explainee are crucial in AI explainability. The explaining mediates these. For more explainable models, the AI system, as the explainer, provides easy-access information to explain its decision-making. However, the role of the explainer is not just to provide raw data or evidence; it also includes analyzing its meaning and presenting this information in a manner that is accessible and comprehensible to the human user. This involves considering the user's background, expertise, and the context in which the AI system is used. Therefore, the explainer's role is not passive; it actively tailors the explanation to meet human requirements. It is of utmost importance for AI systems and human designers to collaborate at every stage of the AI product delivery process, from blueprints to output, to achieve explainability. The explainee, typically the human user, engages with the explanation provided by the AI system. The explainer's role is to interpret, understand, and contextualize the explanation within their domain of expertise. This process is not straightforward, as it requires the human user to apply their

[1] Some may argue that machine reasoning could also be worth exploring. I agree with this point, although it falls outside the scope of the current thesis. Still, I consider this aspect for future in-depth analysis.

knowledge, experience, and judgment to interpret the information provided by the AI system. The explainee's role is critical in ensuring the explaining is understood, actionable, and relevant to the decision-making process.

Sometimes, the role of the explainer is bifurcated between the AI system and the explainability tools. The AI system executes the decision-making process, while the tools act as translators, making these decisions understandable to human users. This two-step explanation process complicates the explainer's role, as it introduces potential discrepancies between the model's actual decision-making process and the tools' representation of it. The explainee's role also becomes more challenging in this context. Human users must understand the explanations provided by these tools and their limitations and potential biases. This requires higher critical thinking and awareness of the methods used for explainability. Furthermore, some models often need iterative feedback between the explainee and the explanation system. This iterative process helps refine and tailor the explanations to the user's needs and understanding. It underscores the dynamic nature of explainability in AI systems, in which explanations are not straightforward or inherently clear.

In summary, the roles of the explainer and explainee are interdependent and collaborative. While the AI system generates the initial data, the human user plays a vital role in diagnosing and applying this information. This collaboration is essential for the effective use and trust of AI systems, particularly in complex domains where decisions have significant implications. The separation between the decision-making process and explanation generation introduces multiple layers of responsibility. It requires a collaborative, iterative approach among the AI system, the explainability tools, and the human users. Understanding these dynamics is crucial for effectively implementing and trusting AI systems in complex, high-stakes domains.

## 3.8 Conclusions

In this chapter, by introducing formal terminology, I propose a novel theoretical framework that provides order and opportunities for improved explanation design in the XAI community. The framework allows for dissecting explanations into evidence (factual data derived from the model) and interpretation (a hypothesized function that describes how the model utilizes the evidence). The explanation results from applying the interpretation to the evidence and is presented to the target user through an explanation interface. These features enable the design of more principled explanations by defining the atomic components and the properties that enable their operation. There are three core properties: *(i)* the explanatory potential for the evidence (*i.e.*, how much of the model the evidence can tell about); *(ii)* the faithfulness of the interpretation (*i.e.*, whether the interpretation is true to the decision-making of the model); *(iii)* the plausibility of the explanation interface (*i.e.*, how much the explanation makes sense to the user and is intelligible). I demonstrate that the theoretical framework can be applied to explanations from various methods that align with the proposed atomic components.

The lesson learned from analyzing explanations within the context of the proposed framework is that humans (both stakeholders and researchers) should be involved in the design of explainability as soon as possible in the AI-powered software design process, especially in sensitive application domains, where a blind application of black-box approaches hampers the right to an explanation. Involving stakeholders enables the proper specification of each component of the theoretical framework of explainability and informs model design. This theoretical formalization is a prerequisite for the methodolo-

gical pipeline introduced next. In particular, the constructs of *evidence*, *interpretation*, and *explanation* form the analytical dimensions used to design and evaluate explainability in subsequent chapters. The following chapter operationalizes these concepts into measurable components within real-world ML pipelines.

# II

# Framing Explainability in High-stakes Domains

# 4

# Evaluating Faithfulness

***This chapter is based on:*** *M. Rizzo et al. 'Evaluating the faithfulness of causality in saliency-based explanations of deep learning models for temporal colour constancy'. In:* Explainable Artificial Intelligence*. Ed. by L. Longo, S. Lapuschkin and C. Seifert. Vol. 2155. Cham: Springer, 2024, pp. 125–142. ISBN: 978-3031637995. DOI: 10.1007/978-3-031-63800-8_7. URL: https://doi.org/10.1007/978-3-031-63800-8_7*

Building upon the theoretical framework established in the previous chapter, we now address another research question of this thesis: *How can explanations be systematically evaluated?* While the preceding discussion laid the conceptual foundations for understanding what constitutes an explanation and how it relates to evidence and interpretation, this chapter operationalizes those ideas within a concrete empirical challenge. Evaluation is not a peripheral task in XAI; it is the cornerstone that determines above all whether explanations genuinely reflect models' reasoning or merely simulate understanding. In fact, faithfulness occupies a privileged position among the dimensions introduced in the *triple frontier of explainability*. If intelligibility concerns the human accessibility of explanations and aligns their ethical and contextual adequacy, faithfulness addresses the epistemic integrity of the explanation itself: does it truthfully represent the mechanisms underlying the model's decision? A plausible but unfaithful explanation is not only scientifically meaningless but also potentially dangerous, as it can foster misplaced confidence in unreliable systems and distort decision-making in high-stakes domains.

This chapter tackles the problem of evaluating faithfulness by examining one of the most widespread and intuitively appealing explanation artifacts for DL models: **saliency**

**maps**. Saliency attribution methods aim to identify which parts of the input most strongly influence the model's output, often visualized as heatmaps. Yet, as Chapter 2 highlighted, such visual explanations frequently conflate correlation with causation, offering plausible narratives that fail to capture the model's genuine reasoning. To address this limitation, we develop a causality-oriented evaluation protocol. The proposed approach tests whether saliency maps encode *causal* dependencies between input features and model predictions, rather than merely statistical associations.

Through this lens, the following sections connect theoretical insight with empirical validation, transforming faithfulness from an abstract desideratum into a measurable property of modern DL explanation techniques.

## 4.1 The Seductive Plausibility of Saliency Maps

Saliency maps — heatmaps that highlight input pixels or regions deemed influential for a model's prediction — have become a staple of XAI, especially, but not only, in Computer Vision (CV). Their appeal is immediate and intuitive: a highlighted lesion on a scan or the bounding region of a detected object appears to answer the question "where is the model looking?" and, by implication, "why did it decide this?". This ease of visual interpretation explains why saliency has been widely adopted by both practitioners and domain experts.

While saliency maps appear to offer direct insight into a model's "reasoning", they in fact blur the distinction between visual intuition and epistemic explanation. An interpretable visualization is not necessarily an explanatory one; it may reveal patterns that humans can perceive intuitively without clarifying the causal structure underlying the model's decision. This distinction is critical for XAI, as visual explainability often relies on cognitive biases — particularly the human tendency to attribute agency and intentionality to systems that display structured patterns. Saliency, therefore, operates as a cognitive interface as much as an analytical tool: it satisfies the feeling of understanding without necessarily improving the fact of it.

Precisely because saliency explanations are visually compelling, they are also dangerously misleading. In practice, such maps are often interpreted as indicators of a model's attention, suggesting that the model "looks" at certain regions more than others. This anthropomorphic reading is seductive but conceptually fragile: a visually plausible heatmap may give the illusion of understanding where none exists. The literature surveyed in Chapter 2 already warned that visual plausibility is often a poor proxy for faithfulness. A saliency map that aligns with human expectation can still be unfaithful — that is, it can suggest causal importance where none exists. A model may learn to rely on subtle confounders (watermarks, imaging artifacts, dataset-specific backgrounds), while a saliency method highlights the "right" region for correlational reasons. As Chapter 5 demonstrates, in high-stakes contexts such as medical diagnosis, this deceptive plausibility can lead to misplaced trust and real-world harm.

The tension between visual plausibility and causal faithfulness reflects a broader methodological gap in evaluating explainability methods. Most existing metrics assess saliency *qualitatively* — through expert inspection or alignment with human attention — rather than by testing whether the highlighted regions are necessary for the model's prediction. As a result, plausibility-based validation tends to reinforce anthropocentric expectations rather than reveal model-internal dependencies. A shift toward *causal evaluation* thus requires not only different tools but also a reconceptualization of what counts as evidence in an explanation.

There are three main issues that create a gap between plausibility and faithfulness in model explainability. First, *correlation can be mistaken for causation*; saliency may emphasize features that are statistically linked to a label rather than the features that actually drive the model's computations. Second, many *saliency methods are highly sensitive to minor input changes or adjustments in model parameters that do not impact predictive performance.* This sensitivity suggests that the saliency maps reflect unstable heuristics instead of robust causal relationships. Finally, *it is possible to manipulate models or input data so that saliency maps convey specific visual narratives*, even when the model's true decision-making process differs substantially from the narrative presented.

These problems motivate the need for a causality-oriented evaluation of saliency: we must determine whether a highlighted region is not merely correlated with, but *causally* relevant to, the model's output.

### 4.1.1 Scope: sequential (video) data and TCC as a testbed

Most foundational saliency research has targeted static images or token-level explanations in NLP. Sequential data, and video in particular, introduce additional complications because temporal interactions (what happened in previous frames) and spatial features (what appears in the current frame) jointly determine predictions. Prior work on attention and its explainability in NLP and CV is informative but inconclusive when ported to temporal audiovisual tasks [34, 83, 193, 166]. Empirical behaviors documented for single-frame models do not automatically generalize to architectures that combine convolutional spatial encoders with temporal recurrence (e.g., CNN+LSTM).

We therefore focus on **TCC** as a representative sequential vision problem. In TCC, the task is to estimate the illuminant color affecting a target frame using information from preceding frames; the application is practical (video color correction) and conceptually suitable for causal testing because temporal dependencies are expected and explainable (the illuminant is a physical property that evolves slowly across frames). Figure 4.1 illustrates the typical CNN+LSTM pipeline used by current SotA methods (e.g., TCCNet [143]).

Sequential data offer a natural laboratory for testing causal explainability. Temporal dependencies imply directionality and contextual conditioning — core elements of causal reasoning. Hence, TCC provides not only an applied benchmark but also a theoretically appropriate setting to test whether saliency captures causal rather than merely statistical relevance.

More broadly, temporal reasoning tasks such as TCC exemplify the next frontier for explainability: models that integrate information across modalities and time are inherently non-local, making the mapping between input and output both delayed and distributed. Traditional saliency approaches, which assume an instantaneous relationship between a single input and its prediction, are poorly suited for such settings. This motivates the development of explanations that trace the propagation of influence over time — a shift from static saliency to dynamic causal attribution.

### 4.1.2 What it means to "explain" in-model attention

To operationalize faithfulness for in-model mechanisms, we make explicit the two elements that constitute an explanation (Chapter 3): the *evidence*, namely the internal signals the network produces to weigh or prioritize features, and the *interpretation*, that is,

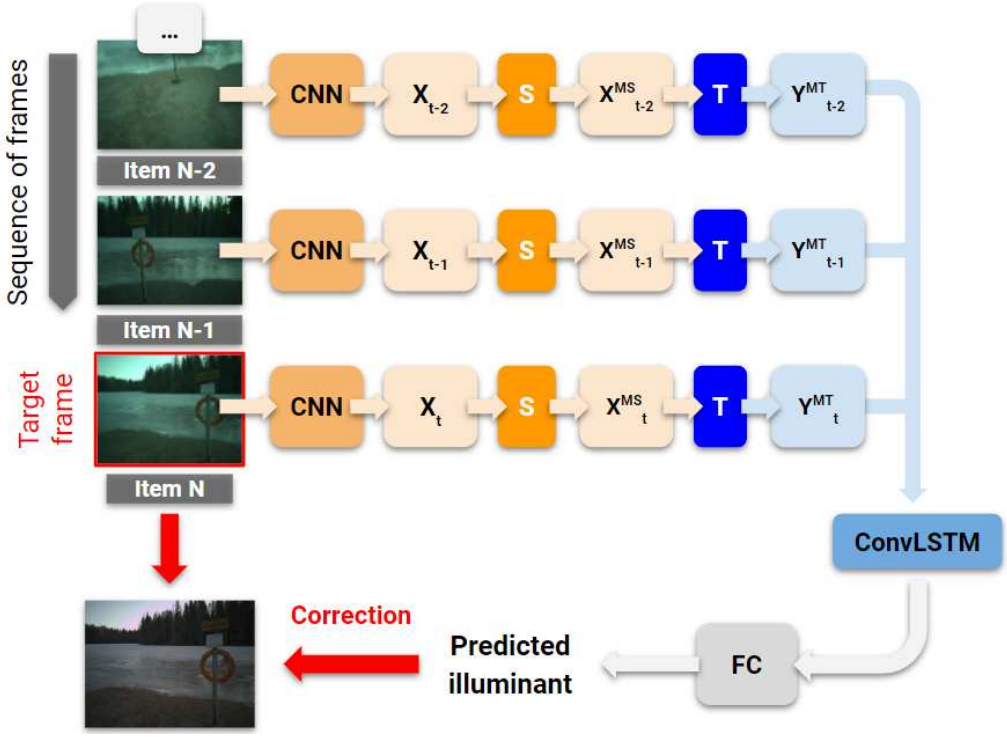



Figure 4.1: Example of CNN+LSTM architecture for the TCC task.

the semantic claim that these weights correspond to causal importance for the model's decision. A faithful explanation requires that this interpretation accurately mirrors the model's internal computational process.

In architectures that employ attention, such as those integrating LSTM-based temporal reasoning, attention weights serve as dynamic selectors over representations. They modulate how much the model "focuses" on particular spatial or temporal components when forming its output. Attention is thus an explicitly relational mechanism, as it defines the relative importance of competing inputs or features. By contrast, the "confidence" weights introduced in early CNN-based color constancy modules [78] are not comparative selectors but multiplicative scaling factors applied to feature channels. They quantify the model's estimated reliability for each feature map, thereby indicating how strongly a channel contributes to the final output. While attention distributes emphasis across features, confidence regulates the amplitude of individual contributions.

We adopt a working definition of a *causal interpretation* that unifies both mechanisms under a common criterion. An evidence signal — whether attention or confidence — is causally interpretable if deliberate interventions that alter highly weighted components induce commensurate changes in the model's output, whereas interventions on low-weight components do not. This definition transforms faithfulness from an intuitive notion of plausibility into an empirically testable property grounded in causal inference.

This intervention-based definition aligns with recent trends in causal explainability, which treat explanations as counterfactual claims rather than descriptive summaries. In this view, a faithful explanation is one that remains valid under hypothetical manipulation of the model or its inputs — an idea consistent with Pearl's structural causal models and with the interventionist theories of explanation in the philosophy of science. By grounding saliency evaluation in causal inference rather than in visual coherence, this approach bridges formal notions of explanation and the empirical practices of deep learning

research.

### 4.1.3 Research design and operational tests

To investigate these claims, we instrument the baseline CNN+LSTM architecture with both attention and confidence mechanisms, applied independently to the spatial (CNN), temporal (LSTM), and joint spatiotemporal components, yielding nine distinct model variants. We first verify that introducing these mechanisms does not materially affect predictive accuracy, ensuring that any differences observed in causal behavior arise from explainability structures rather than performance artifacts. We then conduct a series of controlled intervention experiments, adapted from causal testing frameworks in NLP, to assess the faithfulness of each model's saliency representations within the TCC task.

This evaluation combines two complementary diagnostics. The first is a feature ablation test, in which input regions or features that receive high attention or confidence are selectively perturbed or removed, and the resulting change in the model's output is measured. Faithful causal explanations should produce larger output deviations when influential regions are disrupted. The second diagnostic, counterfactual reweighting, manipulates the internal saliency weights directly—amplifying, suppressing, or nullifying them—while keeping inputs constant. If the model's output changes in accordance with these weight manipulations, the saliency signals can be regarded as causally faithful. Together, these tests move evaluation beyond descriptive alignment toward empirical falsification of causal claims.

Importantly, these tests do not presume that causal faithfulness is an intrinsic property of saliency; rather, they treat it as an empirical hypothesis subject to falsification. This stance introduces a measure of epistemic humility: rather than assuming that visual explanations reveal the model's inner workings, we explicitly test whether they do. Such falsification-based protocols transform explainability research from a rhetorical practice — producing visually pleasing heatmaps — into an experimental science of model interpretation.

### 4.1.4 Contributions and expected gains

This chapter advances explainability research in three ways. First, it extends causal faithfulness diagnostics, originally developed for NLP and static vision models, to sequential video analysis, using TCC as an exemplar domain where temporal dependencies make causal reasoning both tractable and interpretable. Second, it provides a systematic comparison between attention-based and confidence-based saliency mechanisms across spatial, temporal, and combined settings, mapping the conditions under which each form of in-model saliency more reliably reflects causal relevance. Third, it adapts and validates intervention-based evaluation protocols for CNN+LSTM architectures, offering a principled alternative to purely visual inspection and enabling reproducible, quantitative assessments of explainability in temporal vision systems.

The sections that follow detail the model variants and datasets, formalize the intervention protocols, and present both quantitative and qualitative results. These findings are then discussed in light of broader theoretical concerns about faithfulness, intelligibility, and alignment introduced in previous chapters.

## 4.2 Related Work

Prior studies on computational color constancy, which primarily focus on single images, have briefly explored the role of attention in improving accuracy but have not examined explainability in depth [199, 206]. This work also investigates the confidence method, first introduced by Hu et al. [78] for single-frame color constancy, which improves accuracy and suggests potential for explainability.

We assess these in-model saliency methods by the faithfulness of their causal interpretations. Following the framework proposed in Chapter 3, we categorize saliency weights as *evidence*, their causal relation of importance to the model output as *interpretation*, and the highlighted input components as *explanation*. However, evaluations of saliency-based explanations' faithfulness, particularly in video data, are scarce and often lack a clear distinction between evidence, interpretation, and explanation [166, 83, 193].

In this thesis, we extend two tests proposed by Wiegreffe & Pinter [193] to video data. These tests evaluate the faithfulness of attention in NLP tasks by adapting them to assess the causal interpretation of the saliency scores. The first test (WP1, based on the original authors' initials, hereafter) is designed to assess whether attention weights have a meaningful impact on task accuracy. The second test (WP2 from now on) aims to determine whether attention weights embed information about the relationships among input timesteps. Details on these tests are provided in section 4.4.

Before Wiegreffe & Pinter [193], Jain & Wallace [83] proposed two different tests to evaluate the faithfulness of attention in NLP tasks. One test compares with alternative measures of input feature importance, *e.g.* gradient-based measures, assuming that attention-based importance is faithful if the feature importance weights it generates highly correlate with those generated by the other measures. We do not consider this test because it relies on the unverified assumption that the alternative feature importance measures are accurate. The second test examines whether replacing the learned attention weights with different distributions affects model predictions. It is assumed that if this change does not affect prediction, then the weights are not involved in the decision process. Thus, they cannot provide faithful explanations of such a process. This test complements WP1, mentioned above, by verifying the importance of learning weights via the attention mechanism, whereas WP1 suggests that any weights play a role in the decision process.

The subsequent tests conducted by Serrano & Smith [166] aim to understand how well attention weights represent the importance of the encoded input components. This is achieved by setting specific weight values to zero and observing the impact on predictions. Their findings indicate that attention weights are poor indicators of the importance of encoded input components. However, their methodology is limited to plotting trends and lacks a quantitative assessment of model success. These are common issues across the existing evaluations of faithfulness [83, 193, 166], which we address when applying the WP1 and WP2 tests from Wiegreffe & Pinter [193] to the TCC task.

In contrast to these prior studies, this contribution is both methodological and conceptual: we reinterpret saliency evaluation through the lens of causal faithfulness, thereby bridging disparate diagnostic traditions (gradient-based, attention-based, perturbation-based) under a unified epistemic criterion.

## 4.3 Proposed Neural Architectures

To rigorously evaluate saliency faithfulness in TCC, we experimented with nine distinct CNN+LSTM models encompassing three dimensions: Spatial (S), Temporal (T), and spatiotemporal (ST). This study also investigated two saliency types, attention (A) and confidence (C), with particular interest in their potential to enhance explainability. Additionally, we explored a hybrid approach, denoted as CA, which integrates confidence for spatial information and attention for temporal aspects, aligning with their initial design purposes [78, 10].

The CNN+LSTM architecture (depicted in Figure 4.1) includes a spatial saliency module that processes each CNN-encoded frame $X_i$ and learns a mask $MS_i$. The sequence of masked encoded frames, $X^{MS} = X \cdot MS$, is then input into the ConvLSTM, which is equipped with a temporal saliency mechanism that learns a temporal mask, $MT_i$. The output from this process is a series of temporally encoded timesteps $Y_i$ weighted as $Y^{MT} = Y \cdot MT$. After processing through a Fully Connected (FC) layer, this output is utilized for illuminant prediction and subsequent color correction of the last frame in the sequence.

The implementation of attention modules, both spatial and temporal, was adapted from Meng et al. [122]. Spatial attention is learned through a three-layer CNN module that reduces the feature maps to a single channel, using batch normalization and ReLU activations in the first two layers, followed by a Sigmoid activation in the final layer. Temporal attention involves computing the Softmax over the outputs of two feed-forward neural networks, which are jointly trained with the rest of the system. At each timestep, the temporal attention mechanism considers every timestep in the encoded sequence $X_i^{MS}$ and the previous hidden state $H_{t-1}$, resulting in a weighted sum of features from all frames fed into the ConvLSTM.

For confidence, it is spatially oriented, as in its original conceptualization for single-frame computational color constancy [78], learned as an additional channel alongside feature maps, and used to weight the encoded images. Temporal weights are derived by averaging the values of spatial confidence masks, which correlate with the accuracy of single-frame predictions [78].

Both attention and confidence can be visualized through heatmaps (Figure 4.2), providing intuitive insights into the influential input features that affect model predictions. This study builds upon earlier TCC research employing CNN+LSTM models [143] and is a fundamental exploration of in-model saliency in neural networks. Future research will expand this analysis to more complex deep-model designs, such as transformers.

## 4.4 Original Methodology of the Tests

The original test methodology refers to "attention" and is thus reported in these terms, even though this analysis also involves "confidence" saliency. Faithfulness is investigated in terms of the interpretation that saliency scores have a causal relation of importance to the model output.

Test WP1 is designed to assess the role of attention in enhancing the accuracy of deep neural architectures for specific tasks and datasets. It particularly examines whether the attention mechanism contributes to more accurate predictions, a key factor in assessing its faithfulness in a model's decision-making process. The concept of faithfulness here refers to how well the attention mechanism reflects the model's actual computational process in

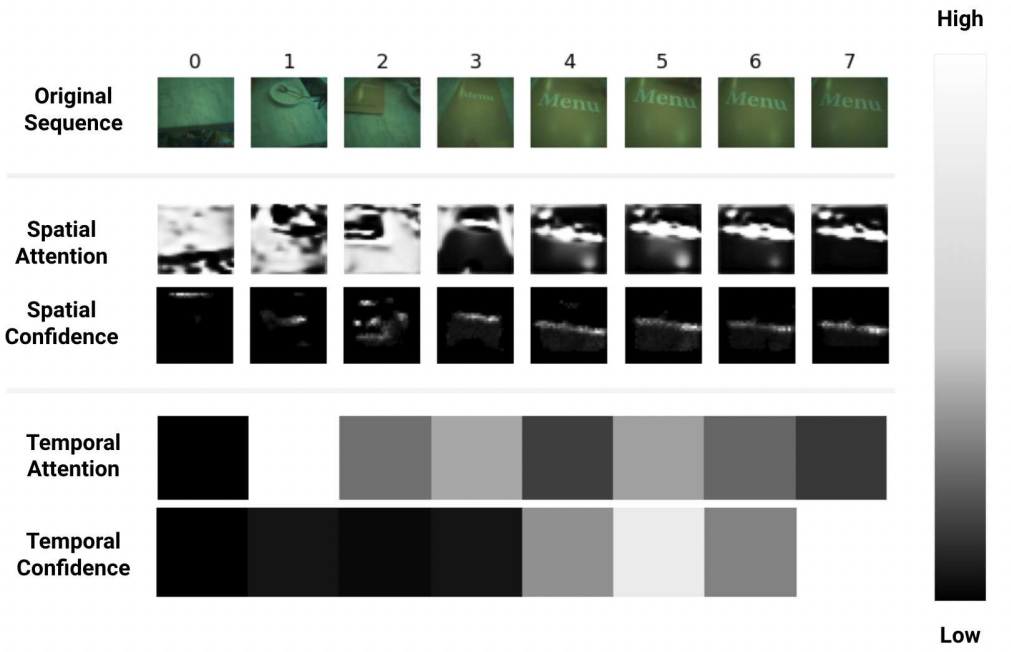



Figure 4.2: Saliency heatmaps for attention and confidence.

making decisions.

In WP1, the performance of a contextual model denoted as $M^C$ is evaluated in two scenarios: (i) using its standard learned saliency weights ($M^C_C$), and (ii) using an alternative version with randomly assigned uniform weights ($M^C_U$). The contextual model, $M^C$, typically includes a recurrent layer, such as an LSTM, to model temporal dependencies among input components, in contrast to a Non-Contextual (NC) model that uses linear layers. The effectiveness of the attention mechanism is validated if $M^C_C$ achieves higher accuracy than $M^C_U$. If $M^C_C$ outperforms $M^C_U$, then attention plays an active role in the model's decision-making, making it a candidate for further investigation regarding its faithfulness. Conversely, if $M^C_C$ does not surpass $M^C_U$, attention may not significantly contribute to the decision-making process, thus questioning its faithfulness. This test establishes a necessary condition for faithfulness, setting the stage for the subsequent WP2 test.

Test WP2 delves deeper into the contextual nature of attention. It investigates whether the saliency weights learned by the attention mechanism encode contextual information about the input components. Contextual information here refers to the understanding of how different parts of the input relate to one another and to their collective impact on the model's output. WP2 tests this by replacing the weights in a Non-Contextual model ($M^{NC}$)—one that does not naturally capture temporal or sequential relationships—with weights learned by a contextual model ($M^C$). The aim is to see if introducing these contextual weights into a non-contextual setting enhances the model's decision-making process, as reflected by improved accuracy. This is measured by comparing the performance of $M^{NC}$ with contextual weights ($M^{NC}_C$) against both its original performance with non-contextual weights ($M^{NC}_{NC}$) and the baseline uniform weights performance from WP1 ($M^C_U$). A positive result in WP2 suggests that the attention mechanism is not merely learning random weights; instead, it captures and transfers valuable contextual information from $M^C$ to $M^{NC}$. This outcome reinforces the role of attention in highlighting crucial parts of the input, leveraging an understanding of the relationships among these

components—i.e., the contextual information. The comparison is made fair by training the linear layers in $M^{NC}$ alongside the other layers, ensuring an equitable basis for assessing accuracy between contextual and non-contextual models.

## 4.5 Method

To bolster the robustness of this work's faithfulness evaluations in TCC, this study incorporated a four-fold cross-validation using diverse training-test splits of the TCC dataset [143]. This methodological choice was driven by the need to balance the sizes of the training and testing samples, given the overall dataset size. Consequently, the findings are reported as mean values and standard deviations across these splits. A significant part of this analysis focused on the Mean Angular Error (MAE), a metric selected to concisely capture the central tendency of angular errors in predicting illuminants (see the supplemental material for additional metrics).

For WP1, this approach compared the performance of models using learned saliency ($M_C^C$) with that of models employing frozen random uniform weights ($M_U^C$). This comparison was conducted using paired t-tests, with p-values adjusted through the Benjamini-Hochberg method to account for multiple comparisons. In the context of WP2, we conducted ANOVA tests with MAE as the dependent variable. Factors in these tests included the dimension of saliency (spatial, temporal, or spatiotemporal), the type of saliency (attention, confidence, or combination), and the nature of weights used in inference (random uniform, contextual, or non-contextual). The results from these ANOVAs were further refined through post-hoc analysis using the Tukey-HSD data method. Additionally, the magnitude of the effects in both t-tests and ANOVAs was quantified using Cohen's d, a statistical measure of effect size.

The proposed methodology also examined the divergence between sets of saliency weights, particularly when models of identical architecture but with different saliency weights were compared. This divergence was measured using the Jensen-Shannon Divergence (sequential data) for temporal saliency distributions and a combination of binary cross-entropy, structural similarity index, and intersection over union for spatial divergence. This combined approach enabled us to evaluate divergence at the pixel, patch, and feature-map levels.

The interplay between saliency weights divergence and model accuracy becomes particularly pertinent when no significant difference in accuracy is observed between models. In such cases, we identified three distinct scenarios: (i) a significant discrepancy in accuracy regardless of saliency weight divergence, (ii) a minimal difference in accuracy accompanied by substantial divergence in saliency weights, and (iii) both minimal differences in accuracy and saliency divergence. In scenarios (i) and (ii), the absence of a notable accuracy difference likely indicates that saliency weights do not play a significant role in the model's decision-making process. In contrast, scenario (iii) necessitates additional investigation to ascertain whether the observed pattern is due to the saliency not being faithful to its intended causal interpretation, potential model overfitting, or a ceiling effect resulting from the simplicity of the task.

## 4.6 Results

In this section, we discuss how the saliency-augmented models compare to the baseline in terms of accuracy. Then, we examine the faithfulness of the generated saliency maps in relation to their causal interpretation. We apply the two selected tests for faithfulness (WP1 and WP2) to three saliency types across three dimensions of a saliency-augmented CNN+LSTM architecture. The saliency types are Confidence (C), Attention (A), and a combination of both (CA). The dimensions are Spatial (S), Temporal (T), and Spatiotemporal (ST).

### 4.6.1 Preliminary Accuracy Investigation

While making a neural model more transparent by modifying its architecture, we would like its accuracy to remain unaltered (or possibly to increase). With this premise in mind, we analyze the impact on the accuracy of augmenting a CNN+LSTM architecture with a saliency mechanism. Specifically, we examine how the nine proposed saliency models compare against a baseline CNN+LSTM that does not incorporate saliency mechanisms in terms of MAE. Despite all models performing worse than the baseline in absolute terms, this trend was not significant under t-tests. The small effect sizes for attention spatiotemporal, attention temporal, and confidence temporal (A-ST, A-T, and C-T) indicate that these models are likely to be equivalent to the baseline in terms of accuracy. More details on these results are available in the supplemental material.

### 4.6.2 Test WP1

Figure 4.3a compares the MAE achieved by models using either random uniformly distributed saliency weights ($M_U^C$) or saliency weights derived from learned model parameters ($M_C^C$). In terms of sheer numbers, we observe that the error achieved by models using random, uniformly distributed weights is always higher than that achieved by models using weights derived from learned model parameters. This trend is particularly pronounced when spatial confidence is incorporated into the evaluation. Running t-tests followed by Benjamini-Hochberg adjustments for multiple comparisons confirms that the trend is statistically significant (p-value $< 0.05$) for each of the examined configurations. Moreover, the corresponding effect sizes are substantial (Cohen's d $> 1$). Thus, all nine of the proposed models pass test WP1. This means that the learned saliency scores convey valuable causal information to the model, enabling accurate predictions. Accuracy is sensitive to manipulation of the saliency distribution.

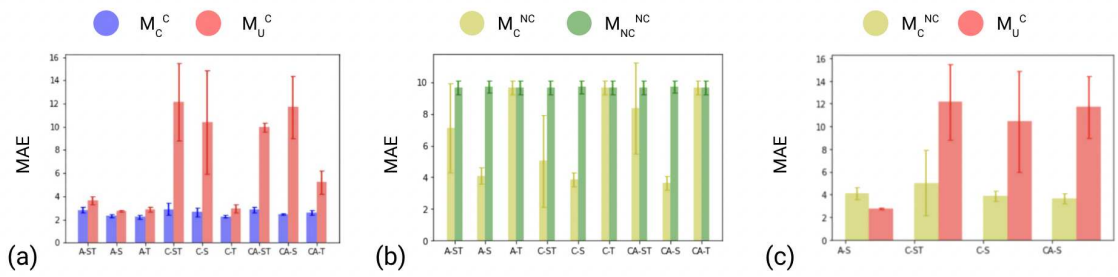



Figure 4.3: Plot of MAE values for Test WP1 (a) and Test WP2, comparison (i) (b) and (ii) (c).

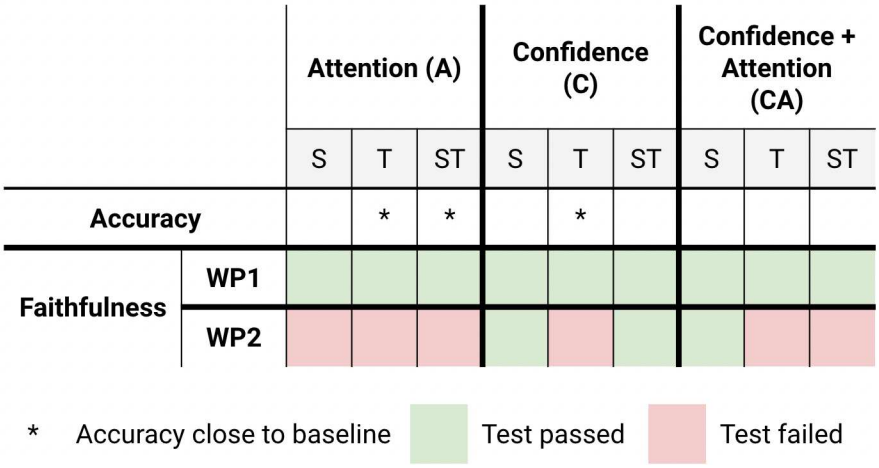



Figure 4.4: Summary table of the results of tests WP1 and WP2.

### 4.6.3 Test WP2

We remark that test WP2 is passed if the MAE achieved by the non-contextual model using saliency weights derived from a contextual model ($M_C^{NC}$) is lower than (i) the MAE achieved by the non-contextual model using weights derived from its learned parameters ($M_{NC}^{NC}$), and (ii) the MAE achieved by the contextual model using frozen uniformly distributed saliency weights ($M_U^C$). We performed these two comparisons for each of the nine proposed saliency models using the contextual model as the reference. Figure 4.3b presents the MAE values for comparison (i) for the considered saliency types and dimensions concerning the type of weights the model uses.

The bar chart shows that (i) holds for spatial attention, spatial confidence, combined spatial attention and confidence, and spatiotemporal (A-S, C-S, CA-S, C-ST). For the other configurations, we need to check if the lack of a significant difference could be due to low divergence in the saliency masks generated by the non-contextual model using learned saliency weights ($M_{NC}^{NC}$) and the non-contextual model using saliency weights imposed from the contextual model ($M_C^{NC}$). We examine the relationships between accuracy and the generated saliency masks for the temporal and spatiotemporal models, interpreting them as discussed in §4.5. For all of the considered models, saliency divergence is high (*i.e.*, $Div_{temp} > 0.7,\ Div_{spat} > 125$), which suggests that the low difference in accuracy is not due to saliency weights being very similar, but instead to them not being involved in the decision-making process of the model in terms of causal interpretation. Therefore, we do not consider these models in (ii).

As shown by the plot in Figure 4.3c, comparison (ii) holds for confidence spatiotemporal, confidence spatial, and confidence and attention combined spatial (C-ST, C-S, CA-S). Thus, these models pass the test WP2. On the other hand, comparison (ii) does not hold for attention spatial (A-S). In this case, the model's contextual architecture has a greater impact on accuracy than the saliency mechanism.

### 4.6.4 Discussion

Figure 4.4 summarizes the results of the accuracy analysis and the two tests for the nine models. First, we note that the temporal dimension is present in all three top-performing model configurations, indicating that this aspect of saliency is crucial for accuracy. However, no model appears faithful to this causal interpretation when using only the temporal

dimension. An intuitive justification for this phenomenon can be observed by noting that temporal saliency tends to focus on a few frames within a sequence. This means that a significant amount of potentially relevant spatial information is discarded. Thus, a model might learn not to actively use temporal saliency in its decision-making process to preserve accuracy.

Second, we observe that spatial confidence is present across all configurations that pass the two assessments, suggesting that this saliency dimension and type support the faithfulness of the causal interpretation. This might be because confidence scores are jointly learned with the other feature maps and are therefore more closely integrated into the model's internal decision-making process. On the other hand, attention configurations never succeed in upholding faithfulness. The crucial difference between attention and confidence is that a separate ad hoc convolutional module learns the former, while the latter is learned as an additional feature map. As a result, the attention model is more complex (*i.e.*, has a more significant number of trainable parameters, $\sim$x3). Attention networks may become sufficiently complex to achieve high accuracy while ignoring saliency in their decision-making. That is, there could be a trade-off between model complexity and the causal interpretation of saliency. A model may be sufficiently complex to solve the task effectively, but introducing additional parameters can lead to a ceiling effect, limiting its performance. In this case, the model could still fulfill the target task without fully leveraging its complexity, regardless of the learned saliency weights.

These empirical observations support the theoretical claim that faithfulness should be regarded as an epistemic, rather than aesthetic, property of explanations. Spatial confidence mechanisms appear to encode evidence more tightly coupled to the model's decision process, achieving greater faithfulness under causal testing. Attention, by contrast, offers higher intelligibility (clearer heatmaps) but weaker epistemic alignment. This dissociation between faithfulness and intelligibility mirrors the tension explored in Chapter 3, underscoring the need for evaluation methods that integrate both dimensions rather than optimizing for one.

## 4.7 Conclusions

In this study, we explored in-model saliency methods, particularly attention and confidence, in the novel context of video-based illuminant estimation, focusing on their faithfulness in influencing model predictions. This assessment covered three dimensions: spatial, temporal, and spatiotemporal. We adapted two tests from previous NLP research to evaluate the causal relationship between saliency scores and model predictions. We enhanced the methodology by incorporating statistical analysis and examining saliency-weight divergence. The findings suggest that spatial and spatiotemporal confidence may be faithful to their causal interpretation, whereas attention models generally fail these tests. This aligns with previous research that has questioned the reliability of attention as an explanatory tool for causal interpretation. On the other hand, the promising results achieved by confidence highlight the importance of integrating in-model saliency to drive faithful causality in explanations. Additionally, this accuracy analysis showed that temporal models tend to perform better.

However, this study has limitations, primarily due to its restriction to a single task and dataset, which challenges the generalizability of the results. Future work will expand these assessments to a broader range of datasets and tasks, including those from the NLP literature, and explore diverse model architectures, such as transformers and different at-

tention methods (*e.g.*, roll-out and flow [1]). We acknowledge the need to establish a clear threshold to distinguish a test's failure due to insufficient divergence in saliency weights. Additionally, we plan to investigate other properties, such as robustness and plausibility. Despite these limitations, this research provides a foundational analysis of the faithfulness of in-model saliency in the TCC task, addressing methodological gaps in previous studies.

This chapter demonstrates that causal evaluation of in-model saliency provides a principled path forward. By grounding empirical analysis in the theoretical framework of faithfulness, it transforms the assessment of explanations from a subjective visual judgment into an evidence-based, interventionist methodology. The next chapter builds on this foundation by exploring how such insights can be scaled and integrated into broader evaluation pipelines. These pipelines also account for intelligibility and alignment, completing the multidimensional evaluation of explainability introduced in this thesis.

# 5

# Assessing the Medical Stakes of Explainable Artificial Intelligence

***This chapter is based on:*** *G. Frasson et al. 'Assessing the value of explainable artificial intelligence for magnetic resonance imaging'. In:* Explainable artificial intelligence. xAI 2025*. Ed. by R. Guidotti, U. Schmid and L. Longo. Vol. 2576. Communications in Computer and Information Science. Cham: Springer, 2026, pp. 320–334.* ISBN: *978-3032083166.* DOI: 10.1007/978-3-032-08317-3_20

The trajectory of this thesis now turns from the conceptual and methodological foundations of explainability to their empirical examination within domains of profound social stakes. In **medicine**, a single misinterpreted or unjustified AI decision can have consequences far beyond just affecting prediction accuracy. In such settings, explainability becomes not merely a desirable feature but a prerequisite for trust and accountability.

Nowhere is the demand for trustworthy AI more acute than in healthcare. As predictive models increasingly assist clinicians in diagnostic and prognostic decision-making, the intelligibility of their outputs becomes essential for their integration into medical practice. This chapter examines both the *application* and the *value* of XAI in this context, emphasizing that what makes an explanation successful is not only its faithfulness to the model's reasoning, but its capacity to contribute to trustworthy and accountable human decision-making.

This chapter begins by addressing our research question: What explanation properties are needed for the effective deployment of AI? Specifically, it investigates this matter within a critical context where explainability intersects with human expertise and institu-

tional responsibility. The case study presented here focuses on the analysis of MRI data, a paradigmatic example of complex input in which model performance is deeply entangled with the clinician's capacity to understand, evaluate, and act on algorithmic outputs.

## 5.1 The Unique Challenge of Clinical Adoption

Where previous chapters have focused on the formal and epistemic dimensions of explanation — its faithfulness — the clinical domain foregrounds a distinct, **human-centered** dimension of value. In this context, an explanation must operate as a *communicative act* between the model and the clinician. Its adequacy is therefore determined not solely by internal faithfulness, but by its pragmatic alignment with the clinician's diagnostic process.

The explanatory object here is not abstract: it must support concrete decisions under uncertainty. For an explanation to be meaningful, it must render the model's reasoning *intelligible*, offering insights that can inform differential diagnosis or guide further investigation. It must also be *aligned*, structured to integrate with the inferential routines of expert radiologists, rather than introducing exotic artifacts. Finally, it must be *trust-calibrating*, allowing the clinician to gauge when reliance on the AI is warranted and when caution is appropriate. MRI offers a particularly stringent test of these demands. Its interpretative complexity — arising from subtle anatomical variations and context-dependent diagnostic cues — renders it an ideal proving ground for assessing whether explainability can meaningfully mediate between model reasoning and clinical understanding.

## 5.2 Assessing the Value of Explainability for MRI

It goes without saying that AI has revolutionized medicine in recent years, driving significant advancements across diverse domains, from drug design and discovery [185, 126] to clinical decision support [171, 85]. In particular, AI has demonstrated remarkable potential as a decision-support tool in medical diagnostics [131, 136]. A study by McKinney et al. [117] reported that an AI system for breast cancer diagnosis, used to interpret mammograms, reduced false positives and false negatives by 5.7% and 9.4%, respectively. Multiple studies have observed that AI systems can outperform human experts in specific diagnostic tasks, thereby enhancing physicians' capabilities through assisted analysis. For instance, Kim et al. [93] demonstrated that an AI system is more sensitive in diagnosing breast cancer than radiologists, effectively identifying early-stage cases. Similarly, Haenssle et al. [67] showed that their AI model achieved superior diagnostic performance for melanoma cases, outperforming most, though not all, dermatologists involved in the study.

Despite these promising results, the widespread adoption of AI in clinical practice is hindered by a critical challenge: physicians are unlikely to trust an algorithm's decision without a clear understanding of its reasoning process. Thus, the topic of XAI is particularly sensitive in medicine, where ethical considerations and regulatory frameworks necessitate accountability and fairness. For example, the European Union's General Data Protection Regulation (GDPR, Article 15) and AI Act grant patients the right to understand how and why decisions affecting them are made. A comprehensive review by Van der Velden et al. [179] examines explainability methods applied to medical imaging across anatomical regions, emphasizing the growing importance of XAI in healthcare.

This study focuses on DL-based analysis of MRI scans to diagnose Distal Myopathys (DMs), a rare Neuromuscular Disease (NMD). Radiological diagnosis of this condition requires significant expertise, as early-stage cases often exhibit subtle tissue alterations that are challenging for less-experienced observers to detect. AI systems can assist radiologists by identifying these patterns and providing supporting evidence for their predictions. This thesis aims to move beyond classification by generating explanations that clarify the rationale for the model's decisions and by investigating their effectiveness. Notably, the existing literature presents several gaps: (1) current explainability methods for MRI-based diagnosis have primarily focused on common diseases with large datasets, with limited attention given to rare conditions such as DMs; (2) existing saliency-based methods often generate noisy, low-resolution explanations that are difficult for clinicians to interpret; and (3) few studies have evaluated the practical clinical relevance of XAI outputs through direct user studies with radiologists. To address these gaps, I introduce two novel explainability techniques tailored to the MRI-based diagnosis of rare neuromuscular disorders: a hierarchical occlusion method and an ensemble explainability strategy. The hierarchical occlusion provides a multiscale view of regional importance by systematically masking image patches at multiple resolutions, thereby improving the localization and clarity of the model's attention. The ensemble explainability strategy aggregates multiple explanation maps to produce more robust and stable outputs, thereby reducing artifacts and enhancing explainability. I benchmark these approaches against SotA methods and conduct a user study with expert radiologists to validate the clinical utility of the resulting explanations. Their feedback assesses the trustworthiness, explainability, and usability of AI-generated explanations, providing critical insights into their potential adoption in real-world medical practice.

While high diagnostic performance is necessary, it is insufficient for clinical adoption without transparency. The theoretical framework in Chapter 3 treats explainability as a *relational* process among the model, its outputs, and the human explainee. For MRI-based diagnosis of rare NMDs, XAI should not only surface predictive features but also *align* with radiologists' reasoning and domain knowledge. This study operationalizes these principles by embedding human-centered evaluation into the design and assessment of explainability.

The following research questions drive the study:

- **RQ1:** How accurately can DL models classify MRI scans for DMs, and what factors influence their misclassification?
- **RQ2:** How do expert radiologists perceive AI-generated explanations and their clinical relevance?
- **RQ3:** How does the radiologist's experience impact the understanding and trust in explainability techniques?
- **RQ4:** What improvements are needed to enhance AI explainability for clinical adoption?

The code for the study is publicly available at https://github.com/matteo-rizzo/xai-for-mri.

## 5.3 Theoretical Framing of the Problem

This empirical study operationalizes the theoretical framework defined in §3.4. This begins by explicitly framing the explanatory process of clinical decision support in terms of *explainer*, *explaining*, and *explainee*. Within this paradigm, the *explainer* is not merely the AI model but a human–machine composite in which domain expertise, model design, and raw computational outputs converge. The *explaining* mediates this relationship: it structures and contextualizes the *evidence* into a form interpretable by the *explainee*, embedding both evidence and its *interpretation* in a coherent, actionable *explanation*. The *explainee*, in turn, is the clinician who engages with the explanation, understands it, and integrates it into their diagnostic reasoning.

In the study, the *evidence* is derived from a DL model trained to classify MRI scans according to the presence of specific pathologies. The *evidence extractor* includes internal mechanisms of the examined methods, based on saliency maps and gradient-based attributions, which generated information about which anatomical regions influenced the model's predictions. The *explaining* process then transformed this evidence into a saliency map overlaid on the MRI scan, thereby conveying insights into the model's decision-making.

Expert radiologists served as both explainees and evaluators. They were tasked with assessing explanations based on four human-centered axes: (i) the *usefulness* of the highlighted area for diagnosis, (ii) the *appropriateness* of the highlighted region's size, (iii) the *ease of interpretation*, and (iv) the perceived diagnostic *reliability*. This framing emphasizes that explanations are both relational and context-dependent. Especially in the clinical context, the practical utility of an explanation depends not only on algorithmic faithfulness but also on the alignment between the explanation and the explainee's diagnostic practices. In this view, XAI plays a dual role: it reveals the model's internal reasoning while simultaneously guiding the clinician's understanding and calibrating trust.

To ensure that the generated explanations are plausible and understandable, the system must be designed to actively consider stakeholder needs. For this medical use case, a preliminary analysis of the radiologists' diagnostic workflow and clinical needs was conducted prior to the evaluation. This initial needs assessment informed the selection of the human-centered evaluation axes (usefulness, appropriateness, ease of interpretation, and reliability). The subsequent user study serves as a post-analysis to assess how well the XAI outputs align with the established preliminary requirements.

## 5.4 Related Work

NMDs comprise a vast and heterogeneous group of pathologies affecting muscles and the nerves that control them. These conditions manifest in childhood and adulthood, presenting significant diagnostic challenges due to their variable clinical features. Diagnosis involves an evaluation of the patient's history and symptoms, supplemented by instrumental examinations, including electromyography, muscle imaging, genetic analyses, and a muscle biopsy.

Recent research has explored the potential of AI to improve diagnostic accuracy for NMDs. Pineros et al. [140] and related work on muscle MRI [144] underscore the utility of AI in this domain. Verdù-Dìaz et al. [183] analyzed patterns of muscle fatty replacement in T1-weighted MRI scans of 976 pelvic and lower limb scans—quantifying fatty infiltration with the Mercuri score and applying a Random Forest classifier—to achieve

an accuracy of 95.7% compared to experts. Yang et al. [201] developed a model for differentiating dystrophinopathies from other muscular diseases using 432 thigh-focused MRI cases. The ResNet50 architecture achieved 91% accuracy, surpassing expert diagnoses that ranged from 80% to 84%. Complementary studies include Felisaz et al. [49], who compared multiple Machine Learning (ML) models for predicting fat fraction and muscle water T2 from MRI texture analysis, and Fabry et al. [47], who employed a 1-Lipschitz neural network on whole-body MRI examinations to distinguish facioscapulohumeral dystrophy from myositis with accuracies between 69% and 77%. While these studies demonstrate the promise of AI in diagnosing NMDs, they also highlight a critical gap: the explainability of model predictions. Only a few works, notably Yang et al. [201], have integrated explainability techniques into their models. This gap motivates the systematic exploration of XAI methods under rare conditions, such as DM. XAI encompasses a range of techniques and methodologies for interpreting the decision-making processes of complex models, particularly deep neural networks, which are often regarded as opaque *black boxes*. Despite their impressive predictive performance, these models lack transparency, which hinders their clinical adoption. The field of XAI remains underdeveloped, lacking a universally accepted taxonomy—a situation partly attributed to divergent definitions of *explainability* and *interpretability*. In this work, I adopt the taxonomy proposed by Linardatos et al. [108], which categorizes methods by dimensions such as model specificity (model-specific vs. model-agnostic) and explanation scope (local vs. global).

Among the well-established approaches for XAI, Class Activation Maps (CAMs) and their extensions have received significant attention. CAMs, introduced by Zhou et al. [209], are post-hoc, local, and model-specific techniques that visualize the discriminative regions used by a CNN to make predictions. By performing global average pooling on the final convolutional feature maps and projecting the resulting weights back onto these maps, CAM highlights the regions most influential to the final decision. However, this method is limited to specific network architectures and only provides explanations from the last convolutional layer. To address these limitations, GradCAM [165] was developed. GradCAM extends CAM by incorporating the gradients of the class score with respect to feature maps from any convolutional layer, thereby generating a class-discriminative localization map through global averaging of these gradients. Nonetheless, GradCAM may struggle to accurately localize multiple instances of an object within an image. GradCAM++ [21] refines this approach by computing a weighted average of the pixel gradients, while HiResCAM [43] further improves explanation fidelity by highlighting only the regions actively contributing to the class score. Comparative analyses indicate that GradCAM tends to produce broader explanations, whereas GradCAM++ and HiResCAM, mainly when applied to architectures such as ResNet50v, may occasionally highlight extraneous background regions—though HiResCAM consistently yields more focused and detailed explanations.

In a contrasting paradigm, SHAP [112] adopts a game-theoretic perspective to assign an importance value to each input feature based on its marginal contribution to a prediction. As a post-hoc, model-agnostic method, SHAP approximates complex models with an additive explanation model that satisfies properties such as local accuracy, missingness, and consistency. Despite the computational challenges inherent in computing exact Shapley values, practical approximations have made SHAP a powerful tool for both local and global explainability.

Another noteworthy approach within XAI is *occlusion*, a sensitivity-analysis method that assesses the impact of masking specific input regions on model predictions. Initially

introduced by Zeiler et al. [204], occlusion systematically masks parts of an input image using a sliding window, thereby identifying regions whose absence leads to a significant decrease in prediction confidence or a change in classification outcome. Although conceptually straightforward, occlusion is computationally intensive, requiring multiple forward passes through the network. Thus, careful selection of parameters such as window size, stride, and occlusion value is crucial; larger windows reduce computational cost at the expense of granularity, while smaller windows offer finer resolution but require more computations. Occlusion is particularly relevant to this thesis, as I propose a novel occlusion algorithm that operates at multiple levels of granularity.

Collectively, these approaches — from CAM-based visualizations to SHAP and occlusion methods—provide a robust foundation for interpreting the decisions of complex DL models. These strategies form the basis of the ensemble method presented in this thesis.

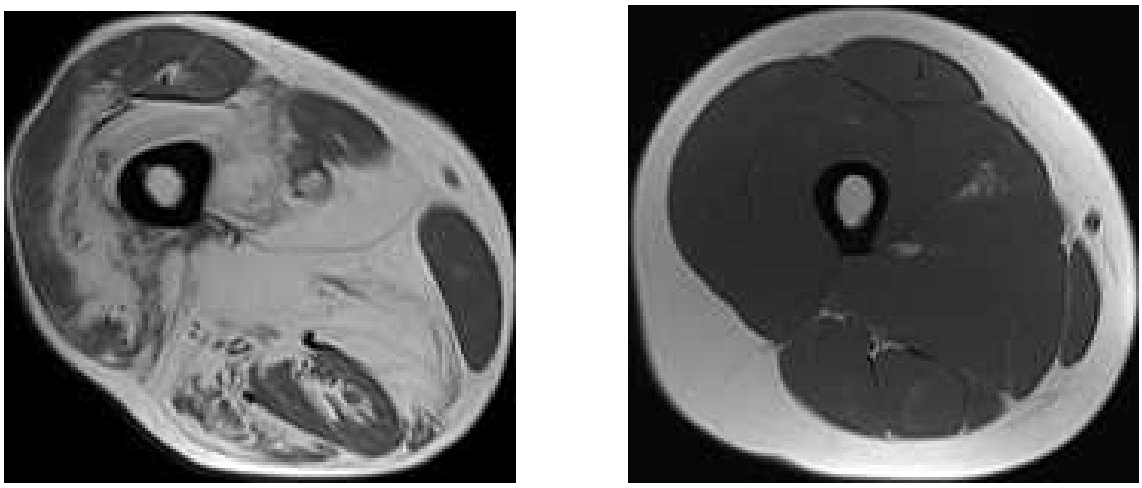
## 5.5 The Use Case: Distal Myopathies

In the context of NMDs, the utility of muscle MRI has already been assessed in the diagnostic work-up and in monitoring the progression of muscle involvement. Given the rarity of these disorders, a thorough clinical, histological, and imaging investigation should be performed, as the clinical heterogeneity and broad genetic spectrum often make a specific molecular diagnosis difficult. In this scenario, muscle MRI is helpful for identifying distinct patterns of muscle involvement. Nonetheless, such clinically and genetically heterogeneous conditions require knowledge of radiological characteristics associated with distinct genetic mutations to improve diagnostic accuracy. Specific patterns are most readily identifiable in patients with mild phenotypes, in which individual muscles are selectively affected. In contrast, extensive and severe muscle involvement, as well as very mild and initial involvement, do not allow for clear pattern detection, even if an expert radiologist can identify them. For this purpose, various AI approaches can be implemented to enhance diagnostic performance, and their explainability will be discussed in this work.

### 5.5.1 Dataset

The proprietary dataset used in this study comprises 529 T1-weighted MR images of the lower limbs, with one image per patient for each side. It comprises seven patients with DM and six healthy controls. To augment the dataset, each image is split into two corresponding to the left and right lower limbs. Although this division could introduce bias, mainly when the disease affects only one side, consultation with an experienced radiologist confirmed that the benefits of a larger dataset outweigh this risk. An example of an affected lower limb is shown in Fig. 5.1a, while an example of a healthy lower limb is presented in Fig. 5.1b. A side-by-side comparison of both cases can be seen in Fig. 5.1.

While the dataset includes only a limited number of patients, this limitation is inherent to the epidemiological rarity of DMs. Currently, there are no large, publicly available MRI datasets specifically tailored to DMs that provide the stringent clinical annotations and imaging consistency required for this rigorous XAI evaluation. Consequently, the scope of this work was deliberately restricted to this proprietary dataset. Although small, it ensures a high-quality, expert-validated ground truth that would be impossible to achieve by aggregating heterogeneous, weakly labeled public data.



(a) (b)

Figure 5.1: Comparison of affected (a) and healthy (b) lower limb MRI scans.

### 5.5.2 Preprocessing

The applied preprocessing pipeline is illustrated in Fig. 5.2. Central to the pipeline is a cropping algorithm that extracts the minor crop that encloses the patient's body from each MRI. Initially, the algorithm enhances the image contrast using Contrast Limited Adaptive Histogram Equalization (CLAHE) [141] to address the uneven brightness and contrast inherent in the original MRI scans. The enhanced image is then binarized by applying a threshold at the mean intensity, which isolates the image's significant regions. Despite producing a clear binary outline of the patient's body, this process can introduce internal holes and noise. To resolve these issues, the outer boundaries of the white regions are detected, and smaller contours are filtered out to retain the two largest contours, which typically correspond to the pelvis and the legs. Subsequently, the rough contours are refined by computing their convex hulls, yielding smoother and more accurate boundaries. The background, which often appears as shades of dark gray rather than pure black in MRI scans, is removed by multiplying the original image by its binary mask, thus preventing any confusion between the background and anatomical structures. Finally, the smallest bounding box enclosing the refined contours is determined using OpenCV's `boundingRect`, and the image is cropped accordingly. Another significant challenge is the heterogeneity in image dimensions, as the scans range from the pelvis to the calf. Since the model requires fixed-size inputs of 224x224 pixels, directly resizing the images is not viable because it could introduce artifacts and distort anatomical proportions. To address this, expert guidance was followed in splitting the pelvis images into left and right sections, as the central pelvis area primarily contains organs rather than muscles. Images smaller than 224x224 pixels are padded to achieve the desired dimensions, while those larger than 224x224 pixels are segmented into 224x224 tiles. If an image exceeds the required size in one dimension only, it is divided into two tiles; if it exceeds the required size in both dimensions, it is partitioned into four tiles. This approach minimizes the number of generated tiles, thereby reducing the risk of bias from mislabeling healthy regions in patients with the disease. The significant overlap between the tiles further ensures that critical border features are preserved. Table 5.1 displays the dataset structure after completing these preprocessing operations.

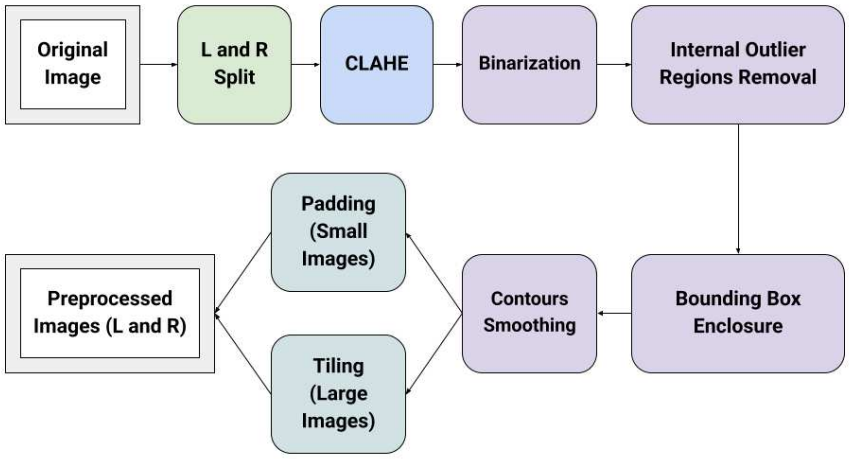



Figure 5.2: Workflow of the preprocessing pipeline.

| | Before preprocessing | After preprocessing | Number of tiles |
|---|---|---|---|
| Healthy | 202 | 404 | 438 |
| Affected | 327 | 654 | 969 |
| **Total** | **529** | **1058** | **1407** |

Table 5.1: Dataset structure after preprocessing.

## 5.6 Models

Due to the limited dataset size, a transfer learning approach was adopted. ResNet-50 was chosen for its strong performance in a similar application reported in [201]. Although both studies address a binary classification task, prior work focused on distinguishing between two diseases, whereas this thesis aims to distinguish between healthy and affected individuals. ResNet18—the lightest Residual network available in PyTorch—was also experimented with. Given the dataset's size, this choice maintained consistency within the model family while reducing complexity and mitigating the risk of overfitting.

Several modifications were made to adapt the pre-trained models to the specific task of this thesis. First, the input layer was adjusted because the PyTorch pre-trained models are designed for 3-channel images, while this thesis's MRI images are in grayscale (1-channel). To incorporate the pre-trained weights appropriately, they were summed across the channels, following the intuition that for an RGB image with equal channel values, $R \cdot w_0 + G \cdot w_1 + B \cdot w_2$ simplifies to $R \cdot (w_0 + w_1 + w_2)$. Next, the global average pooling layer was removed from the network architecture. This change was necessary because techniques such as GradCAM and HiResCAM converge to CAM when applied to networks without global average pooling on the last convolutional feature maps. Accordingly, the architectures described in [43] were modified by replacing the global average pooling with an additional convolutional layer. Finally, the output classifier was replaced. Since the original pre-trained models were configured to predict 1,000 classes (trained

on ImageNet), the final layer was replaced to enable binary classification. The resulting modified models are ResNet18v and ResNet50v, with "v" denoting the variant.

Before training, the dataset was partitioned into training and test sets using a patient-based split to prevent data leakage. This strategy ensured that the data from a single patient did not appear in both sets, thereby preserving the integrity of the evaluation process. However, given the limited number of patients, the test set contained only one individual per class. To address the potential sensitivity of the performance metrics to this selection, an experienced radiologist recommended including a healthy patient with a higher body fat percentage in the test set, thereby challenging the model and providing a conservative lower bound for evaluation.

Due to class imbalance in the dataset, ROSE oversampling was applied to the training set to equalize class counts [119]. Furthermore, the feature-extraction layers were frozen, and only the network's classifier layer was trained using a cross-validation framework with early stopping to mitigate overfitting. Since the number of patients was limited, it was not feasible to reserve a separate validation set with one healthy and one affected individual; instead, each patient was treated as a separate fold in the cross-validation process. During training, data augmentation techniques provided by PyTorch—such as random brightness and contrast modifications—were applied to artificially increase the size and diversity of the training dataset. These augmentations, chosen in consultation with a domain expert, were designed to reflect the natural variations typically encountered in MRI images, thereby improving the model's generalization capability.

## 5.7 Proposed Methods

### 5.7.1 Hierarchical Occlusion

As previously described, *occlusion* is a practical yet computationally intensive sensitivity analysis technique, particularly when a high level of detail is desired. Moreover, selecting appropriate parameters—such as window size, stride, and occlusion value—is nontrivial, as they often require extensive tuning and may not be universally optimal across inputs. To address these challenges, a hierarchical occlusion algorithm was developed. The central idea is to start with relatively large occlusion windows and progressively reduce their size, thereby balancing computational cost against the desired level of granularity in the analysis. The initial concept was to leverage an existing occlusion function, such as the `Occlusion` class from Captum [97]. However, this implementation exhibited two significant limitations. First, it does not allow selection of an alternative metric, as it defaults to the difference in the model's output. Second, it lacks the flexibility to apply occlusion to a designated subarea of the image, a requirement for the hierarchical approach. Consequently, a custom design and implementation were pursued.

The proposed hierarchical occlusion algorithm performs occlusion at multiple levels of granularity. Still, it restricts the refinement process to those windows that, at a coarser level, induce a change in the network's output. Several strategies were explored in the development process. Inspired by Captum, an initial approach employed the difference in the model's raw output before and after occlusion as the metric. However, combining results across different levels proved problematic because each level inherently possesses a distinct value range. Larger windows tend to produce more pronounced differences than smaller windows, thereby complicating direct comparisons. Using the difference in probability—bound in the interval $[0, 1]$—appeared to be a more interpretable alternative,

but similar issues in merging results across varying granularities persisted. Smaller windows naturally yield more minor differences and may erroneously be interpreted as less significant. Furthermore, establishing a universal threshold for further refinement is challenging, given that each image exhibits its own range of output differences. To overcome these issues, an alternative metric is needed that identifies windows that actually cause changes in the network's classification. This approach is more stringent, as it ignores minor variations in the output and focuses solely on occlusion windows that result in a class switch. By assigning a binary value (1 for windows that induce a change in the predicted class and 0 for those that do not), the results from different levels of granularity can be aggregated by summation. The outcome is a composite map that indicates, at varying levels of detail, the regions whose occlusion alters the network's prediction.

The algorithm initially computes occlusions using larger windows across the entire image to contain the computational cost. It then refines the analysis by applying occlusion with progressively smaller windows exclusively to those regions where the initial occlusions caused a change in the network's output. Alg. 1 provides a high-level algorithm description. Several parameters are critical in the proposed implementation. I provide the classification model, the input image, and the target class for occlusion. The initial window size and stride are chosen based on the input dimensions (in this thesis's case, 224x224 pixels), and they are halved at each successive level of granularity until a predefined minimum window size is reached. The occlusion window is filled with a specified value (zero in this thesis's implementation).

Selecting optimal parameters is challenging due to the heterogeneous nature of the dataset, which contains images of body parts with varying sizes, and the need to balance computational efficiency with the quality of the resulting occlusion maps. For instance, given that the network input is 224x224 pixels, the initial parameters were set to a window size of 56 and a stride of 28, with a minimum window size of 7. These settings are designed to capture relevant details without being so fine-grained that they fail to cover larger regions of interest. In practice, the hierarchical algorithm frequently yielded void outputs, meaning that none of the occlusion windows produced a change in the network's prediction. This phenomenon was particularly notable in images of pelves and thighs, and to a lesser extent in images of calves and knees. In response, the window size was increased for images that initially produced void outputs. This change reduced computational overhead by avoiding unnecessary recalculations across all images while preserving finer occlusion maps when available.

**Algorithm 1:** Hierarchical Occlusion

**Input:** Model, image, target class, initial window size, stride, minimum window size, occlusion value

**Output:** The final hierarchical map

$hierarchical\_map, areas \leftarrow$
compute occlusion at level $n$ over the entire image **while** $areas \neq \emptyset$ **do**
  $single\_map, areas \leftarrow$
  compute occlusion at level $n-1$ restricted to regions in `areas`
  $hierarchical\_map \leftarrow hierarchical\_map + single\_map \;\; n \leftarrow n-1$
**end**
**return** $hierarchical\_map$;

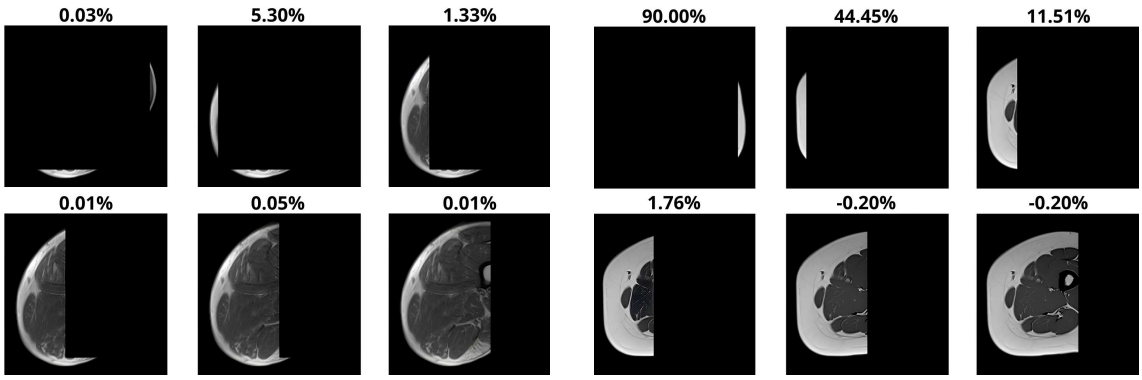
Table 5.2 summarizes the results obtained using the hierarchical occlusion methods. As expected, increasing the window size generally reduces the occurrence of void outputs.

However, an increase in window size does not necessarily correlate with increased importance, as many new activations may result from occluding a large portion of the patient's body. An additional experiment was conducted with an exaggeratedly large window size of 200. Although a larger window is more likely to affect network predictions, Table 5.2 indicates that, particularly for the *affected* class, many images still do not exhibit a response to occlusion. This observation is counterintuitive, as occluding regions in an affected image would be expected to more readily switch the prediction to *healthy*, whereas the converse should be more difficult.

A classical, non-hierarchical occlusion method was implemented to further investigate this phenomenon, using the difference in probability as the metric. Figures 5.3a and 5.3b display the results of the initial occlusion windows for images classified as *affected* and *healthy*, respectively, using ResNet18. The numerical values annotated on the images represent the percentage difference between the original and occluded probabilities. A positive difference indicates a decrease in the network's confidence, whereas a negative value indicates an increase. Notably, the initial occlusions in Fig. 5.3a remained classified as *affected* with high confidence despite the occluded regions corresponding to healthy fat. A similar pattern was observed for the *healthy* images in Fig. 5.3b. This suggests that the network may misinterpret subcutaneous fat as infiltrated fat when analyzed in isolation, potentially indicating an inherent bias toward predicting the *affected* class.

Table 5.2: Occlusion results across different occlusion window sizes.

| **Model** | **Prediction** | **Occlusion Window Size** | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | **56** | | **86** | | **112** | | **200** | |
| | | Void | Not Void | Void | Not Void | Void | Not Void | Void | Not Void |
| ResNet18v | Affected | 109 | 62 | 93 | 78 | 84 | 87 | 81 | 90 |
| | Healthy | 17 | 39 | 7 | 49 | 0 | 56 | 0 | 56 |
| ResNet50v | Affected | 102 | 67 | 88 | 81 | 88 | 81 | 55 | 114 |
| | Healthy | 22 | 36 | 12 | 46 | 6 | 52 | 0 | 58 |




(a) (b)

Figure 5.3: Occlusion windows comparison for affected (a) and healthy (b) MRI scans.

### 5.7.2 Ensemble of Explainability Methods

Given that no single XAI technique consistently outperforms the others, an ensemble approach is adopted to integrate multiple explainability methods. By aggregating the outputs of diverse models, the ensemble capitalizes on the strengths of each technique while compensating for their limitations, yielding more robust and reliable explanations. In this work, the ensemble is constructed by aggregating the non-zero heat maps produced by the various explainability methods, excluding the void occlusion maps. Preliminary experiments combined the outputs of all explainability techniques; however, GradCAM was ultimately excluded from the final ensemble because its tendency to produce broader activation regions was found to dilute the more focused insights provided by the other methods. The proposed ensemble strategy, therefore, comprises GradCAM++, HiResCAM, SHAP, and Hierarchical Occlusion. Before integration, heatmaps are preprocessed to reduce noise and enhance explainability. Rather than operating at the pixel level, the heatmaps are partitioned into 7x7 square blocks, with each block assigned the average of its constituent pixels. Negative values are removed to focus exclusively on areas that contribute positively to the model's prediction. Finally, the data are normalized to the interval $[0, 1]$, ensuring that the outputs from different techniques are on a comparable scale. Three ensemble strategies were investigated, each with a different degree of restrictiveness. Fig. 5.4 compares the base explainability methods and this thesis's three proposed ensemble strategies. The first strategy computes the average of the heatmaps and selects regions where the average exceeds 0.5, a procedure analogous to majority voting; this approach tends to yield broader areas of evidence. The second strategy is based on an intersection approach: heatmaps are first filtered to retain only activation values above 0.2, and the ensemble is then defined as the common regions across all methods. While emphasizing regions with unanimous support, this intersection approach may exclude significant areas according to all but one method. The third strategy focuses on saliency by considering only pixels with values above 0.7 and aggregating those selected by at least $n - 1$ of the $n$ methods. This more selective approach highlights smaller, more precise regions of interest. This study adopted a threshold of 0.7 to achieve a stringent ensemble that emphasizes only the most salient areas.

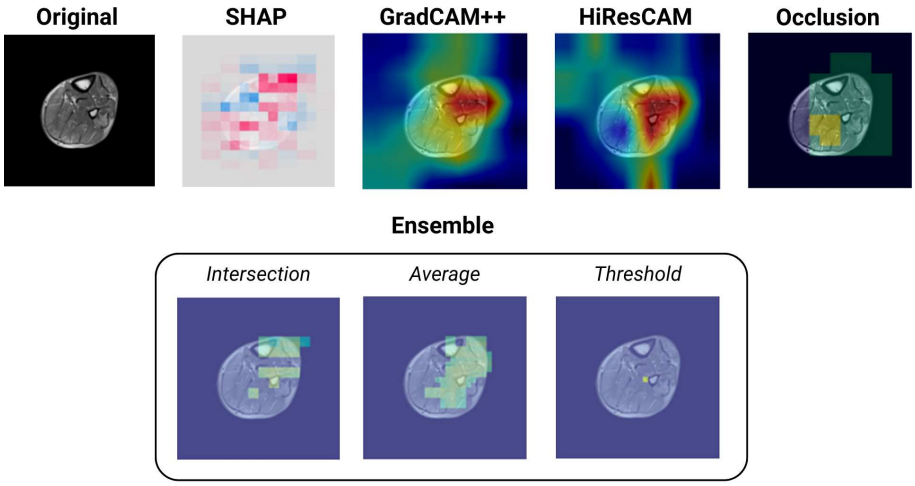



Figure 5.4: Comparison of base explainability methods and proposed ensemble strategies.

The hierarchical occlusion and ensemble strategies presented here extend traditional

XAI methods by explicitly considering both multiscale feature relevance and robustness across techniques. From a theoretical standpoint, these contributions operationalize the concept of *explaining* in §5.3, translating raw evidence into interpretable, clinically meaningful information. Moreover, by involving expert radiologists in the evaluation, the study situates explainability within a real-world decision-making context and directly addresses research questions RQ2–RQ4. This integration of method development and user-centered assessment exemplifies a human-AI collaboration paradigm in which explanations are validated not only for algorithmic accuracy but also for epistemic value.

## 5.8 Results

### 5.8.1 Model Accuracy

The performance of the AI models was evaluated using standard classification metrics, including accuracy, precision, recall, and F1 Score. As summarized in Table 5.3, both ResNet18v and ResNet50v demonstrated strong classification performance, achieving accuracies close to 90%. The high recall scores indicate that both models reliably identified all positive class instances, while precision remained competitive, resulting in robust F1 scores.

| Model | Accuracy | Precision | Recall | F1-score |
|---|---|---|---|---|
| ResNet18v | 88.55% | 84.80% | 100% | 91.77% |
| ResNet50v | 89.43% | 85.80% | 100% | 92.36% |

Table 5.3: Performance metrics for the final models evaluated on the test set.

Analysis of the confusion matrices (Fig. 5.5) reveals that both networks consistently identified all instances of the positive class, while misclassifications predominantly occurred within the *healthy* class. A closer inspection suggests that images misclassified as *affected* often exhibited prominent subcutaneous fat, which may have contributed to the incorrect classification. This indicates that the model may rely on fat distribution patterns as a proxy for pathology. This unintended bias could be addressed by refining preprocessing techniques, such as explicit muscle segmentation or feature calibration.

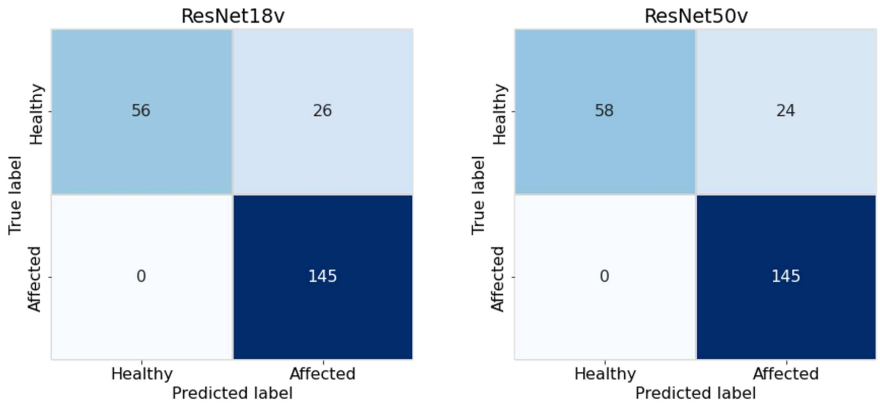



Figure 5.5: Confusion matrices for ResNet18v and ResNet50v.

### 5.8.2 Explainability

Beyond quantitative performance, this study investigated the explainability and clinical relevance of AI-generated explanations through a structured evaluation. The study involved 14 participants: 7 board-certified radiologists, 6 radiology residents, and 1 experienced specialist in neuromuscular imaging. To better contextualize the results, participant characteristics were recorded, including age (median: 35 years), gender (9 male, 5 female), and years of specialized education (median: 6). Exploring the relationship between these characteristics and users' cognitive abilities is crucial in this context. For instance, visual processing strategies and cognitive load differ significantly between novices and experts; residents often rely on high-cognitive-load spatial scanning, whereas experienced specialists utilize rapid, heuristic-based pattern recognition.

#### 5.8.2.1 Diagnostic Accuracy and Observer Performance

Table 5.4 summarizes the accuracy of the radiologists compared to the most experienced observer (Observer G). The average diagnostic accuracy was 80%, with notable variability among residents. While some observers closely aligned with the network's predictions, others misclassified most cases. This variability suggests that some instances are inherently ambiguous, even for human experts, highlighting the potential for AI assistance in diagnostic workflows. Despite the high recall of the AI models, misclassified instances were observed by both the network and human observers. Analysis of confusion matrices (Fig. 5.6) reveals that ambiguous cases lacked conclusive explainability outputs, which may explain neutral or negative ratings regarding their reliability. This finding underscores the subjectivity inherent in XAI techniques and the challenge of ensuring trust in AI-generated explanations.

| **Physician** | **Experience (years)** | **Accuracy** |
|---|---|---|
| Observer A | 1 | 60% |
| Observer B | 1 | 90% |
| Observer C | 2 | 60% |
| Observer D | 2 | 80% |
| Observer E | 2 | 100% |
| Observer F | 3 | 90% |
| Observer G | 9 | 100% |

Table 5.4: Diagnostic accuracy of radiologists on the test subset.

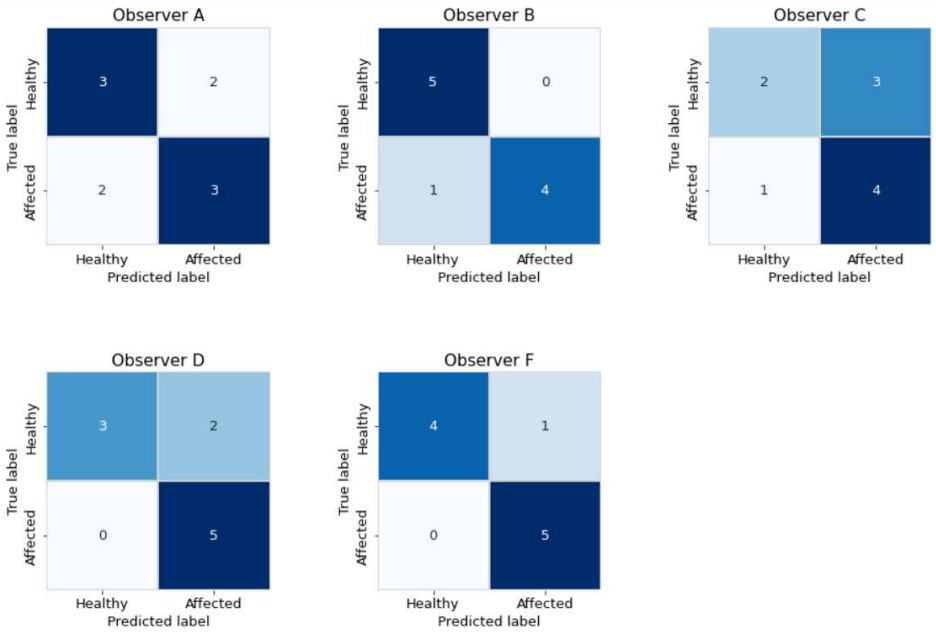



Figure 5.6: Confusion matrices for the observers who committed classification errors.

#### 5.8.2.2 Methods Evaluation

Fig. 5.7 illustrates the votes for different explainability techniques. No method emerged as the clear favorite, as all received at least three votes. GradCAM and SHAP were among the most frequently selected methods. However, a deeper examination of individual preferences shows that Observer F strongly favored SHAP, while GradCAM was chosen by only three out of seven observers. Notably, experienced radiologists did not favor these methods but slightly preferred GradCAM++. This discrepancy suggests that experience influences how explainability methods are interpreted, with expert users valuing refined, localized attributions over broader, attention-spanning visualizations. The evaluations highlight significant variability in the perceived effectiveness of different explainability techniques. Some observers consistently selected GradCAM, whereas those who preferred ensemble methods tended to exclude it. Ensemble-based approaches, although receiving fewer overall votes, were regarded by the experienced radiologist as producing more diagnostically relevant explanations. This preference divergence underscores the importance of tailoring XAI methods to different user expertise levels. Further analysis of the explainability evaluations is presented in Fig. 5.8, illustrating the distribution of observer votes for each image. The chart highlights the high variability in observer preferences, ranging from a maximum agreement of 42% (three out of seven observers) for images 1, 9, and 10, to complete disagreement for image 3, where each radiologist selected a different explainability method. This lack of consensus underscores the subjective nature of explainability assessments and the need for more tailored, adaptable XAI techniques.

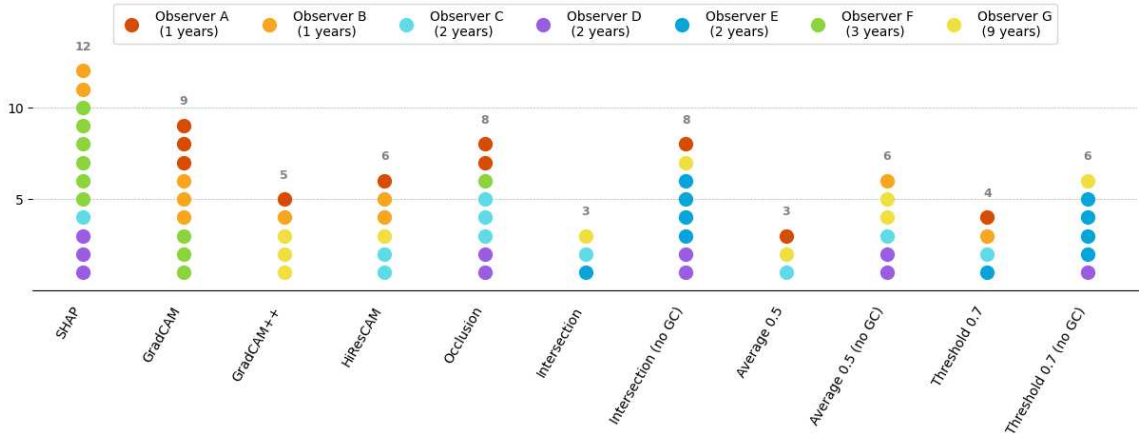



Figure 5.7: Preferences for explainability methods across dataset. GC stands for GradCAM.

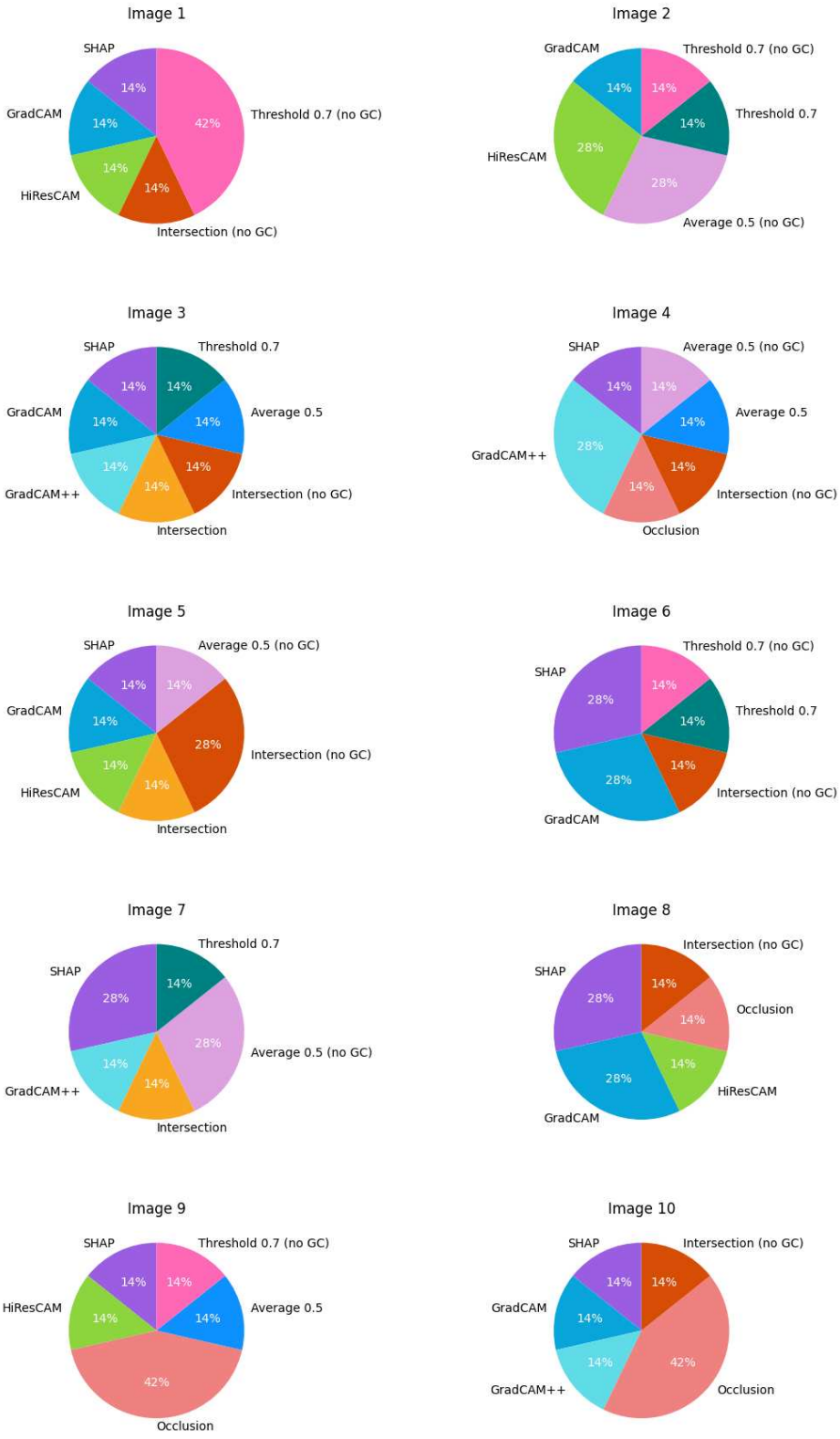



Figure 5.8: Observer preferences for explainability methods across different images.

#### 5.8.2.3 Ratings and Observer Preferences

To further analyze the perceptions of explainability techniques, I examined the distribution of scores across four evaluation criteria: (i) *Usefulness of the highlighted area for diagnosis*, (ii) *Appropriateness of the highlighted region size*, (iii) *Ease of interpretation*, and (iv) *Perceived diagnostic reliability*. Fig. 5.9 presents the score distributions. The median ratings were predominantly neutral or low, indicating limited confidence in the AI-generated explanations. Notably, SHAP and the 'Average 0.5' ensemble received the highest median ratings (4) for diagnostic usefulness. In contrast, occlusion-based methods received lower scores, likely because they occasionally failed to generate clear explanations. For perceived diagnostic reliability, most methods received scores concentrated in the lower range, except for the SHAP and 'Average 0.5' ensemble, which exhibited symmetric distributions around neutral values. This suggests overall hesitation to fully trust the AI-generated explanations. During the evaluation, there were five cases where the AI correctly suggested a diagnosis to an observer who initially misclassified the input. However, in none of these cases did the observer change their decision after reviewing the AI explanations, reinforcing the challenge of aligning AI explainability with clinical reasoning.

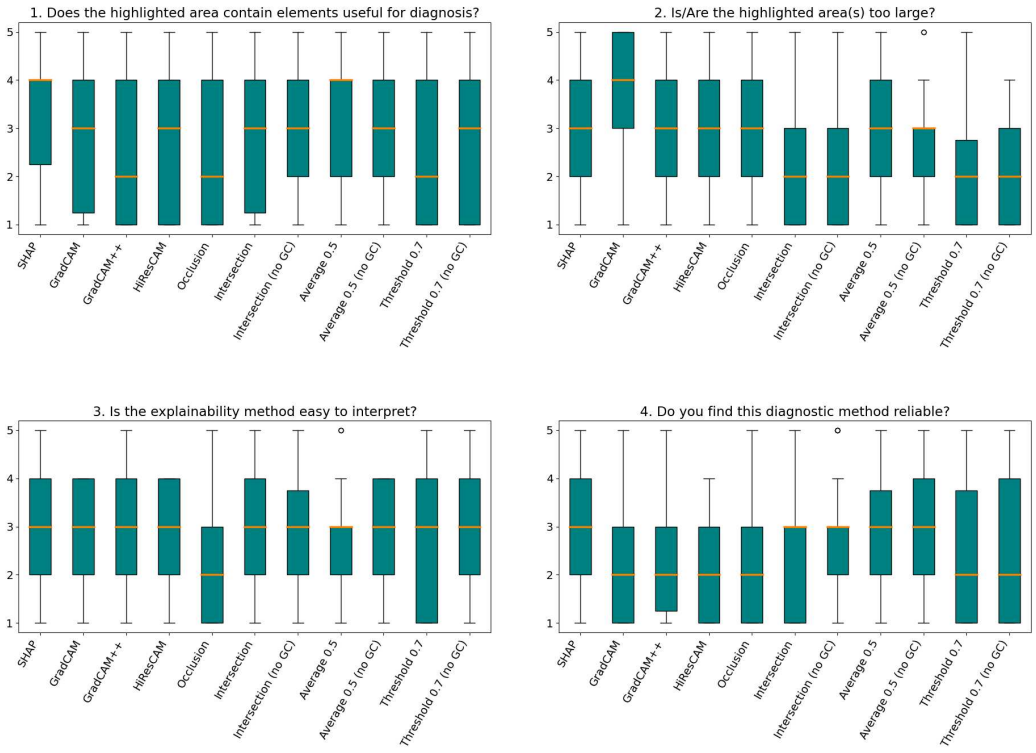



Figure 5.9: Distributions of observer ratings (1 to 5) across different explainability methods. GC stands for GradCAM.

#### 5.8.2.4 Comparison of Individual vs. Ensemble Methods

I grouped the evaluations into Individual and Ensemble methods to assess the effectiveness of ensemble approaches. Fig. 5.10 shows the proportion of observer preferences

for each category. Although Individual methods received slightly more votes overall, experienced radiologists slightly preferred ensemble techniques. Fig. 5.11 displays score distributions for the two groups. Individual methods exhibited greater variability, with a skew towards lower values, whereas Ensemble methods showed a more stable, neutral distribution. Interestingly, although observers preferred Individual methods, they rated Ensemble methods higher for perceived reliability. This suggests that, although ensemble techniques are less familiar, they may yield more clinically meaningful insights.

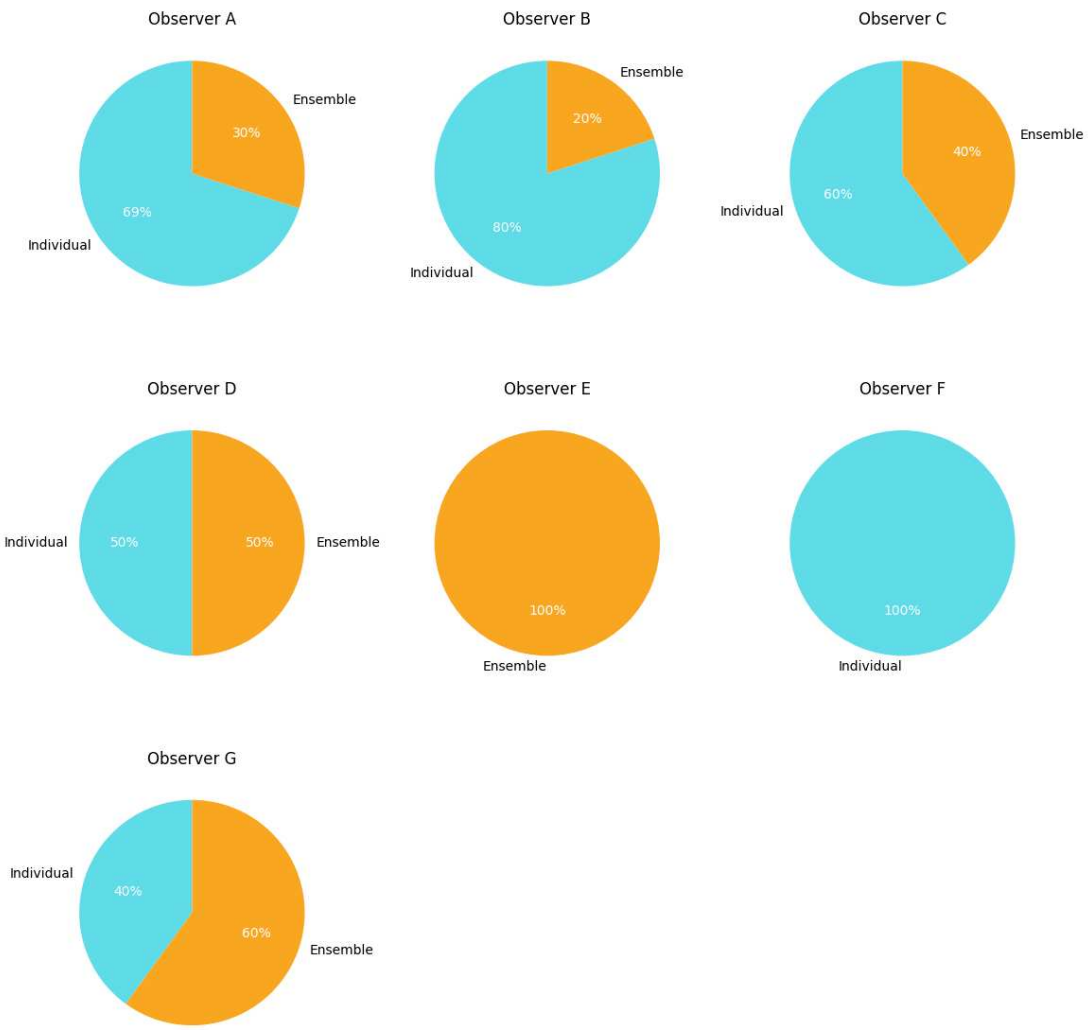



Figure 5.10: Proportional distribution of observer preferences.

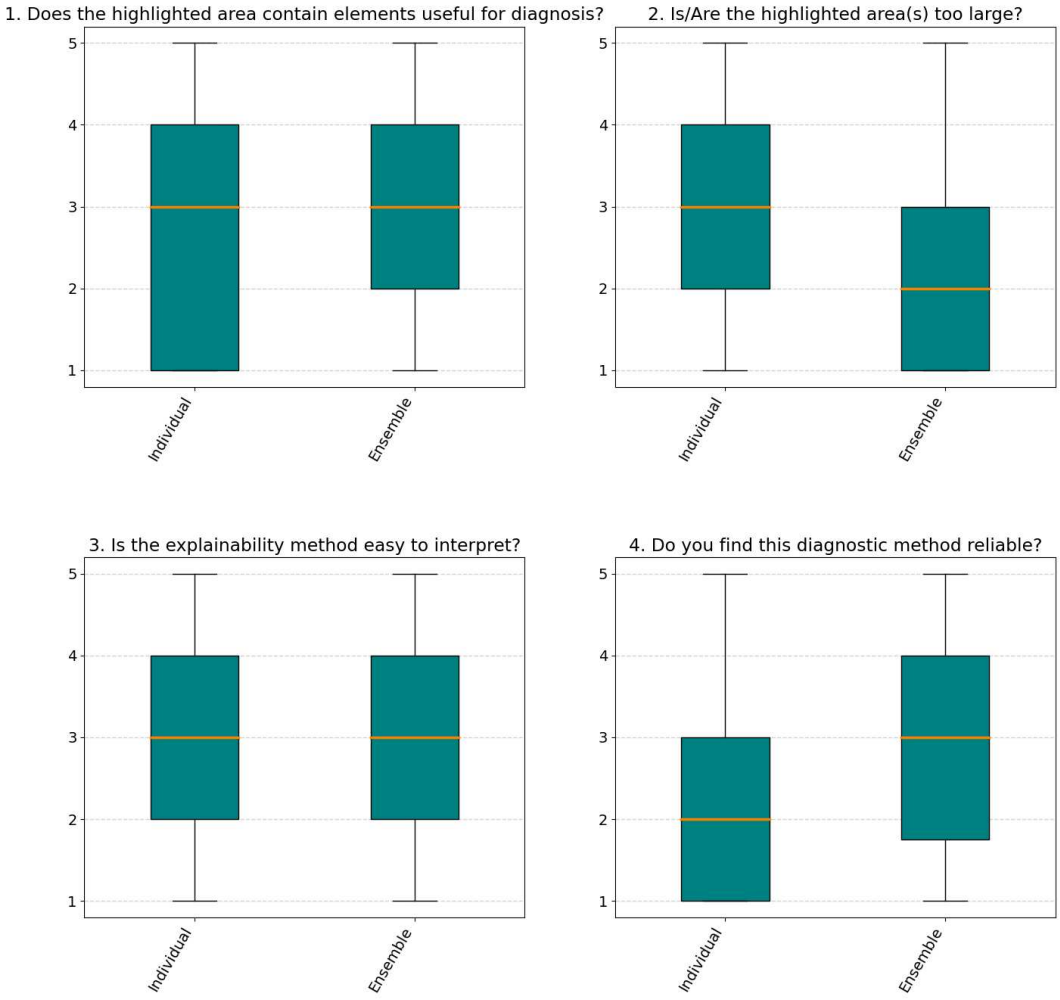



Figure 5.11: Score distributions for Individual and Ensemble explainability methods.

### 5.8.3 Discussion

The evaluation of ResNet18v and ResNet50v confirms that DL architectures can achieve high diagnostic accuracy for rare NMDs, even under limited and heterogeneous data. While overall performance metrics are strong, a detailed examination of misclassifications exposes a subtle yet important limitation: the models occasionally anchor their predictions on features not strictly tied to pathology, such as patterns of subcutaneous fat. This finding is critical for the thesis, as it underscores the latent risk of spurious correlations in medical imaging and highlights the gap between raw performance and clinical explainability. In other words, high accuracy alone does not guarantee clinically meaningful learning; the models' internal representations may encode proxies that are effective for prediction but misaligned with domain knowledge.

The structured observer study further contextualizes this limitation, exposing that explainability is not only a computational challenge but also a human-centric one. Radiology residents and experienced specialists interpreted the same outputs differently: novices gravitated toward visually salient, broad maps, whereas experts favored localized, refined

attributions, often provided by ensemble methods. This divergence underscores a key thesis insight: explainability is inherently relational. Consequently, evaluating XAI methods requires situating them within the end user's expertise and decision-making strategies, rather than treating them as universally explainable artifacts.

A second, intertwined insight arises from the observed disconnect between AI-corrected diagnoses and human adjustments to decisions. Even when the network identified cases misclassified by physicians, the accompanying explanations rarely influenced clinical judgment. This highlights another crucial thesis argument: explainability is necessary but insufficient for trust. Trust emerges not merely from transparency, but from alignment between AI representations and human reasoning. For AI to act as a meaningful diagnostic collaborator, explanations must resonate with clinicians' cognitive and perceptual expectations, integrating both anatomical priors and task-specific relevance.

From a methodological standpoint, the comparative evaluation of individual versus ensemble explainability techniques provides an additional layer of interpretation. While individual methods were more intuitively appealing to less experienced observers, ensemble approaches—though less visually prominent—were rated higher for perceived reliability by expert users. This suggests that precision, consistency, and context-sensitivity may be more important than visual salience in high-stakes medical decision-making. The broader implication is that XAI frameworks in clinical contexts must strike a balance between explainability and epistemic rigor, tailoring the presentation of evidence to both the level of expertise and the diagnostic context.

Finally, these results illuminate the thesis-wide tension between model performance, explanation fidelity, and clinical utility. Even the most accurate model is only as useful as the interpretive framework through which it is understood. This reinforces a central argument of this dissertation: in medical AI, success is not measured solely by metrics such as accuracy or recall, but by the extent to which algorithmic insights can be trusted and operationalized by human experts.

## 5.9 Limitations

While this study provides valuable insights into the diagnostic performance and explainability of AI models for DMs, several limitations must be acknowledged. First, while the proprietary dataset is highly curated and representative of this specific rare condition, its limited size may not fully capture the broader diversity of real-world MRI scans. Extending the experiments to larger datasets was not viable due to the absence of publicly available MRI databases for DMs. Variability in imaging protocols, scanner models, and patient demographics could influence model generalizability. Future studies should validate these findings on larger, more heterogeneous datasets to ensure robustness across clinical settings. Second, the structured evaluation of explainability methods involved a limited number of radiologists, with only one experienced specialist. While this provided valuable insights into how expertise influences the interpretation of AI explanations, a broader sample of radiologists with varying levels of experience would be necessary to draw more generalizable conclusions. Additionally, inter-observer variability suggests that user preferences for explainability methods may be highly individualized, underscoring the need for adaptive, customizable XAI frameworks. Another limitation is the reliance on retrospective image assessments rather than on real-time clinical workflows. The physicians reviewed AI-generated explanations in a controlled experimental setting, which may not fully reflect how these methods would be used in actual diagnostic practice. Future stud-

ies should investigate the impact of AI explanations in prospective settings where radiologists integrate them into routine clinical decision-making. Furthermore, while the study examined various explainability techniques, it did not explore the full spectrum of available XAI methods or assess how combinations of techniques could enhance explainability. Given the variability in observer preferences, future research should investigate hybrid approaches that dynamically combine multiple explanation strategies to better align with radiologists' diagnostic reasoning. Finally, the study primarily focused on explainability and diagnostic performance, without considering the computational efficiency and real-time feasibility of the proposed methods. Some explainability techniques, particularly ensemble approaches, may be computationally expensive, which could potentially limit their practical deployment in clinical environments. Future work should explore optimization strategies to balance explainability quality with computational constraints, ensuring seamless integration into medical imaging workflows.

## 5.10 Conclusions

This chapter demonstrates that DL can provide effective diagnostic support for rare neuromuscular disorders, while also highlighting critical caveats that inform both theory and practice. High model accuracy does not guarantee clinical relevance, as neural networks may exploit statistical proxies that diverge from pathophysiological reasoning. Recognizing and mitigating such biases requires integrating domain knowledge directly into model design and preprocessing. The study further reveals that explainability is not a monolithic concept; it is contingent on user expertise, cognition, and interpretive strategies. Less-experienced radiologists preferred salient, broad explanations, whereas experts valued localized, ensemble-based outputs. This supports a central thesis contention: *explainability must be adaptive, not standardized*. Effective AI explanations are those that engage the user's reasoning process, highlighting relevant features with sufficient precision and contextual alignment to be operationally meaningful.

A notable finding is the trust gap between AI outputs and clinician behavior. Even when AI corrected misclassifications, human observers often disregarded the explanations, indicating that transparency alone is insufficient for adoption. Building trust requires explanations that are not only technically faithful but also diagnostically resonant — essentially, explanations must speak the language of clinical reasoning.

Taken together, these findings underscore three interrelated lessons for the thesis: (i) *robust diagnostic AI requires careful calibration* between predictive performance and clinically grounded feature selection; (ii) *explainability must be user-centered*, modulating granularity and focus to match expertise; and (iii) *human-AI collaboration hinges on trust*, which emerges from alignment between algorithmic rationale and human cognitive models. The next chapter will add further depth to the collaborative role of explanations with users, particularly framing generative explanations as an act of mutual knowledge building.

# 6

# Assessing the Security Stakes of Explainable Artificial Intelligence

***This chapter is based on:*** *Matteo Rizzo et al. 'Advanced Large Language Models Prompting Strategies for Reentrancy Classification and Explanation in Smart Contracts'. In:* Blockchain Technology and Emerging Applications*. Ed. by William Knottenbelt et al. Cham: Springer Nature Switzerland, 2026, pp. 37–56. ISBN: 978-3-032-12335-0*

In security-critical domains, the demand for trustworthy and actionable explanations is absolute. Errors in these environments are rarely inconsequential, as a single flawed inference can propagate through financial systems, critical infrastructure, or autonomous agents with irreversible consequences. The opacity of modern AI models, particularly under adversarial conditions, transforms ambiguity in a model's reasoning into a viable attack surface. Conversely, explanations grounded in verifiable logic can serve as a proactive defense, accelerating vulnerability detection, informing threat modeling, and enabling reasoned intervention over reactive mitigation.

This chapter examines the security implications of explainability through the lens of LLMs. These models represent a fundamental shift in AI, moving beyond predictive modeling toward synthetic reasoning, where models generate not only outputs but also their own justifications. This "generative turn" profoundly alters the nature of XAI. Explanations are no longer static reports attempting to approximate a model's internal state; they are dynamic discourses co-produced through interaction. The explanation is not merely a window into a pre-existing decision process but a new act of reasoning in itself.

To ground this analysis, blockchain security serves as an instructive case study. *Smart*

*contracts*—self-executing programs that enforce agreements on decentralized platforms—operate in immutable, adversarial, and financially critical environments where correctness is non-negotiable. An error in contract logic can result in an irreversible economic loss in an instant, making the explanation of vulnerabilities an urgent and pragmatic imperative. Unlike perceptual tasks such as medical image analysis, reasoning about smart contracts involves symbolic structures, control flow, and inter-contract dependencies. Vulnerabilities emerge from complex logical interactions and are actively sought by adversaries.

Among these, the *reentrancy* vulnerability offers a paradigmatic example. It arises when an external call within a contract allows an attacker to re-enter a function before its initial execution completes, enabling repeated, unauthorized actions, such as fund withdrawals. The challenge is twofold: to detect such vulnerabilities automatically and to articulate the underlying logic of the flaw in a manner that developers can comprehend and act upon.

## 6.1 The Faithfulness Dilemma in LLM-Generated Security Explanations

This study employs LLMs for both vulnerability detection and explanation generation, interrogating their emergent role as instruments for high-stakes reasoning. The dual objective is to (i) *identify smart contracts susceptible to reentrancy* and (ii) *produce natural-language explanations* to guide developers toward effective remediation. The analysis operates entirely at the prompting level, examining how linguistic conditioning—including query design, contextual grounding, and structural cues—influences the resulting explanations. The focus is not on probing the model's internal representations but on examining how explainability is elicited through language.

This approach highlights a central challenge of LLM-based XAI: the divergence between plausibility and faithfulness. LLMs excel at generating coherent, contextually appropriate, and highly plausible narratives. The generated explanations in this study proved actionable and pragmatically useful, enabling developers to trace vulnerable code paths, reconstruct malicious execution sequences, and identify the logical dependencies that facilitate reentrancy exploits. However, their utility does not guarantee their faithfulness. An explanation is faithful only if it accurately reflects the causal chain of operations within the model that produced the initial vulnerability assessment. For LLMs, this causal chain is largely inaccessible and may not even resemble human-like logical deduction. The model could reach a correct conclusion via statistical pattern matching and then generate a logically sound explanation as a separate, post-hoc rationalization. This explanation, while helpful, would not be a faithful account of the model's actual "reasoning."

This tension—between pragmatic utility and verifiable faithfulness—crystallizes the core security risk. An unfaithful explanation can serve as a vector for misdirection. It may lead a developer to conclude that a system is secure (or flawed) for the wrong reasons, thereby misallocating attention or implementing ineffective fixes. The very act of generating an explanation becomes a performative one; its primary function is to produce comprehension, not necessarily to reveal computational truth. This validates a central claim of this thesis: the value of explainability is measured by its capacity to mediate between artificial computation and human sense-making, but in security contexts, this mediation must be audited for faithfulness. In high-stakes domains, we must negotiate the balance between human comprehension and the unverified, and perhaps unverifiable,

logic of the underlying model.

## 6.2 Advanced Prompting Strategies for Detecting and Explaining Reentrancy

Smart contracts are at the heart of the decentralized application ecosystem, yet their security remains a pressing concern. The infamous 2016 DAO hack, caused by a reentrancy vulnerability, stands as a sobering example of the financial and reputational damage that a single bug can inflict. Despite years of research and tool development, vulnerabilities such as reentrancy persist, exposing the limitations of conventional security analysis. Static analysis tools rely heavily on rule-based logic, which often leads to a flood of false positives and limited adaptability. Dynamic tools, while more precise in execution, are constrained by path coverage and the generation of test inputs. Both approaches struggle to explain their findings in a way that is actionable and understandable to developers, often reducing their output to little more than warning flags without context.

Another major limitation of traditional tools is their reliance on fixed definitions of vulnerability. In the case of reentrancy, for instance, different tools employ varying interpretations of the concept. Slither, for example, distinguishes between four categories (Ether-stealing, non-Ether-stealing, benign, and event-reordering reentrancy), whereas other analyzers use coarser or incompatible classifications. This lack of consensus on what constitutes reentrant and what does not limits the adaptability of these tools when new attack patterns emerge or when nuanced interpretations of vulnerabilities are required. Modifying a traditional analyzer to accommodate a novel reentrancy pattern would entail substantial architectural changes and a major re-formalization of the underlying theoretical framework.

LLMs offer a radically different paradigm. Trained on massive corpora of code and text, these models can recognize patterns, reason about logic, and even generate natural-language explanations for their outputs—all without being explicitly programmed for the task.

This work tackles that challenge head-on. Rather than relying on naïve prompting, I investigate how structured reasoning and code-aware retrieval can enhance the accuracy and reliability of LLMs in the context of smart contract security. I dissect how providing models with examples that reflect the actual structure of the code under analysis, rather than just its textual surface, can sharpen their diagnostic capabilities. I also study how guiding the model's reasoning through carefully designed prompting strategies can reduce logical errors and produce explanations that are both interpretable and verifiable [150]. Finally, I explore how combining these two approaches—reasoning guidance and structural context—can yield results that surpass the sum of their parts, moving closer to trustworthy, expert-level auditing powered by LLMs.

Experiments conducted in this work on a curated benchmark of real-world, verified Solidity contracts demonstrate that these strategies not only improve classification accuracy but also foster explainability and reduce hallucination. The result is a system that not only detects reentrancy vulnerabilities but also explains them in transparent, correct, and verifiable ways.

The contribution of this work follows:

- **Structural Equality for RAG:** I propose a novel few-shot RAG strategy that retrieves examples based on structural similarity between programs, measured by

comparing either the Abstract Syntax Tree (AST) or the Control Flow Graph (CFG).

- **Expert-Crafted Chain-of-Thought (CoT):** I design a rigorous CoT prompting template, informed by domain expertise, to enforce a reasoning process that mirrors the logic followed by human security auditors.
- **Strategy Evaluation:** I combine structured CoT prompting with this work's RAG pipeline to develop a hybrid approach that integrates logical reasoning with relevant code context, delivering both accuracy and explainability.
- **Hallucination Mitigation:** I demonstrate empirically that this work's hybrid strategy significantly reduces LLM hallucination, marking a step toward safe and reliable LLM-based tools for smart contract auditing.

The code repository for this work is available for reproducibility at: `https://github.com/matteo-rizzo/advanced-llm-prompting-for-reentrancy`.

## 6.3 Background

This section provides an overview of the foundational concepts for understanding this work. I begin by providing an informal definition of reentrancy as a vulnerability in Ethereum smart contracts, followed by a brief description of analysis techniques based on formal methods traditionally used to address it. I finally introduce LLMs and the advanced prompting strategies—CoT and RAG—that form the basis of the proposed solution in this work.

A smart contract is an immutable, self-executing program stored on a blockchain (in this work's case, Ethereum). While smart contracts are generally immutable—meaning their code cannot be altered once deployed—this immutability can be circumvented through design patterns such as proxy contracts, which have gained increasing popularity in recent years. Proxy contracts separate logic from storage, allowing upgrades to the contract's logic while preserving state. Despite this flexibility, pre-deployment security auditing remains crucial. Among the most studied vulnerabilities is *reentrancy*, especially in Ethereum. It occurs when a contract makes an external call to another contract, potentially malicious, before finalizing its state-changing logic. Suppose the callee contract makes a recursive call back into the original function before it completes, allowing it to repeatedly execute a portion of the code, such as a withdrawal function, thereby draining the contract of its funds. The canonical flawed pattern, known as the "call-before-update" pattern, is illustrated by the following Solidity snippet:

```
function withdraw(uint amount) public {
    // 1. Check if the user has enough balance
    require(balances[msg.sender] >= amount);

    // 2. Make an external call BEFORE updating the balance
    //    (VULNERABLE to re-entrancy)
    (bool success, ) = msg.sender.call{value: amount}("");
    require(success, "Transfer failed.");

    // 3. Update the user's balance (too late)
```

```
        balances[msg.sender] -= amount;
    }
```

To secure implementations, the *Checks-Effects-Interactions (CEI)* pattern has been canonized by the Solidity community, suggesting that programmers perform all state changes (*Effects*) before making external calls (*Interactions*). Such a pattern is only a good practice, though, and two problems may arise: first, programmers may simply ignore it and still produce exploitable code; secondly, subtle scenarios or complex contracts may produce behaviors—hence, exhibit potential exploits—that are difficult to predict and analyze.

### 6.3.1 Formal Methods for Code Analysis

The primary methods for automated smart contract auditing have traditionally been static and dynamic analysis [38]. Tools for *static* analysis inspect the smart contract's source code or bytecode without executing it. They employ techniques such as control-flow graph analysis, symbolic execution, and syntax-directed pattern matching to identify potential vulnerabilities in accordance with a predefined set of rules [113, 48]. While effective at identifying many common anti-patterns, static analysis often suffers from a high rate of false positives and can miss complex or novel vulnerabilities not covered by its rule set [113].

Their counterpart is a tool that performs *dynamic* analysis by executing the smart contract in a simulated environment to observe its behavior [63]. By sending a series of transactions to the contract, these tools (often referred to as fuzzers) explore different execution paths to uncover vulnerabilities that only manifest at runtime. However, their primary limitation is achieving complete path coverage; complex contracts with many possible states make it computationally infeasible to explore every possible execution trace, leading to potential false negatives [63]. A common weakness across both paradigms is their limited explanatory power, often leaving developers to decipher cryptic warnings [38].

While formal methods provide rigorous guarantees, they remain brittle in the face of modern smart contract complexity — motivating the exploration of learning-based and, more recently, language-model-based approaches.

### 6.3.2 Large Language Models for Code Analysis

LLMs are DL (DL) models pre-trained on vast corpora of text and code. This pre-training endows them with a robust, generalized understanding of language, syntax, and logical patterns. When applied to source code, LLMs can perform a variety of "code intelligence" tasks, such as code completion, translation, and bug detection, often with impressive *zero-shot* (no examples) or *few-shot* (a few examples) capabilities [24]. Their generative nature allows them to not only classify code but also to articulate the reasoning behind their decisions in natural language, a key advantage over traditional tools.

To steer LLM behavior toward more accurate and reliable outputs, several advanced prompting techniques have been developed. This work focuses on two of the most prominent: *CoT* and *RAG*. The former is a form of prompting technique designed to elicit more complex reasoning from LLMs. Instead of asking for a direct answer, a CoT prompt encourages the model to generate a series of intermediate, logical steps that lead to the conclusion [192]. By externalizing the reasoning process, the model is less likely to make intuitive leaps and more likely to follow a coherent path, which has been shown to significantly improve performance on arithmetic, commonsense, and symbolic reasoning tasks.

The generated chain of thought also provides a transparent window into the model's "thinking," making its output more interpretable and easier to debug. The latter, RAG, is a framework that grounds an LLM's output in external, verifiable knowledge, thereby reducing hallucinations and improving factual accuracy [103]. A standard RAG pipeline works in two stages. First, given a user query, a *retriever* module searches a knowledge base to find relevant information. Second, the retrieved data is concatenated with the original query to form an expanded prompt, which is then fed to the LLM. This provides the model with relevant, in-session context, encouraging it to base its generated response on the provided evidence rather than relying solely on its internal, parametric knowledge.

Despite their promise, LLMs are not inherently equipped to reason about control-flow-sensitive vulnerabilities, such as reentrancy. Their reliance on surface-level token patterns and distributional similarity can lead to the overlooking of edge cases or to spurious reasoning. Moreover, while advanced prompting strategies improve explainability, they do not guarantee adherence to formal security principles. These limitations frame a critical research gap: designing LLM workflows that integrate domain-specific knowledge, structural program representations, and verifiable reasoning steps to achieve both accuracy and trustworthy explanations.

## 6.4 Related Work

Reentrancy is one of the most critical and extensively studied vulnerabilities in Ethereum smart contracts. Detection techniques have evolved from traditional static and symbolic analyzers to ML (ML) models and, more recently, LLMs. I review these approaches across three categories — static and symbolic analysis, ML methods, and LLM-based systems — and conclude with a discussion of structured prompting for explainable, verifiable reasoning.

### 6.4.1 Static Analysis and Symbolic Techniques

A wide range of static and symbolic tools has been proposed for auditing smart contracts. Early systems such as Oyente [113] pioneered the use of symbolic execution, control-flow analysis, and pattern-based vulnerability detection. While influential, many of these tools are no longer maintained or compatible with recent Solidity versions, particularly those released after 0.8.x. In this work, I focus on three widely adopted tools that are still actively maintained and support modern Solidity: Slither [48], Mythril [31], and Confuzzius [176]. Although not of recent origin, these analyzers remain relevant due to continued development and their distinct analytical strategies. Slither is a static analyzer that utilizes an SSA-based intermediate representation (SlithIR) and incorporates several detectors targeting various forms of reentrancy. Mythril employs symbolic execution and SMT-based solving to identify unsafe low-level calls that lack proper state updates. Confuzzius combines symbolic execution with evolutionary fuzzing to flag suspicious traces involving storage access before and after external calls.

### 6.4.2 Learning-Based Detectors

ML approaches aim to address limitations of static analysis, such as high false positives and rigid pattern matching. Traditional ML and DL techniques—including LSTMs,

GNNs, and Transformer-based models such as CodeBERT—have been applied to vulnerability classification, yielding promising results [149]. These models are often trained on SmartBugs Wild [45], a large but weakly labeled dataset, which is commonly annotated using the very static analyzers they aim to replace. This feedback loop has led to inconsistent definitions of reentrancy and poor support for modern contracts.

Due to dataset imbalance and outdated syntax, it remains difficult to evaluate or generalize such models [146]. The lack of a standardized benchmark hinders meaningful comparisons across architectures, particularly between DL and custom-designed models. Furthermore, most approaches are designed to detect a wide range of vulnerabilities, often without a specific focus on reentrancy, and frequently omit class-specific performance metrics, such as the per-class F1 score.

### 6.4.3 Large Language Models

Recent interest has turned to LLMs for contract vulnerability detection, particularly due to their dual capabilities in code understanding and natural language generation [105]. Single-stage prompts enable zero-shot classification, but models such as GPT-4 still produce high false-positive rates and hallucinated reasoning [77]. Moreover, LLMs performance is highly task-dependent and often brittle when asked to reason over code transformations. Multi-stage LLM architectures attempt to improve precision. GPTLens [77], for instance, separates vulnerability generation and critique using an adversarial "Auditor-Critic" loop. However, even these advanced techniques struggle with reentrancy, often due to the use of poor-quality labels and subtle semantic overlaps between patterns.

This work's contribution lies in the design of principled prompting strategies that integrate domain knowledge into LLM workflows. While generic CoT prompting elicits step-by-step reasoning [192], it does not ensure correctness or completeness in high-stakes domains, such as smart contract auditing. I introduce a procedural CoT framework derived from the CEI threat model for reentrancy, which breaks down the reasoning process into verifiable steps, such as identifying external calls and validating the order of state updates. This approach aligns with recent structured prompting work [104], but applies it in a security-specific setting where correctness is critical.

To improve factual grounding, I incorporate a structurally-aware Retrieval-Augmented Generation (RAG) mechanism [203]. While prior RAG methods rely on semantic similarity, this can mislead LLMs in logic-sensitive tasks. Instead, I compute structural similarity using classic program analysis tools; specifically, CFGs and ASTs extracted with Slither and compared using the Weisfeiler-Lehman kernel. While graph-based similarity has been used for clone detection, its use as a retrieval filter for RAG in smart contract analysis is novel and effective. This hybrid approach—symbolically structured CoT reasoning combined with retrieval guided by code structure—represents a neuro-symbolic direction for vulnerability detection. By grounding explanations in verifiable logic and controlling evidence via structural similarity, I aim to mitigate the opacity and unreliability that have plagued prior ML and LLM-based methods.

By integrating structural program representations into the retrieval and reasoning pipeline, this work's approach situates itself at the intersection of neuro-symbolic AI and software security. Unlike prior work that treats LLM outputs as final, the proposed methodology enforces procedural consistency aligned with formal definitions of vulnerabilities. This positions the thesis not merely as a performance study but as a contribution to the theoretical framework for explainable and verifiable AI in security-critical code analysis.

## 6.5 Methodology

The proposed methodological approach first establishes performance baselines across several paradigms. Then it assesses a series of prompting strategies engineered to enhance the LLM's logical fidelity and factual grounding. All experiments are anchored by a manually validated benchmark dataset and are evaluated using a protocol that measures both predictive accuracy and the qualitative utility of the generated explanations. A consistent set of formal principles for identifying reentrancy (e.g., CEI pattern adherence, valid reentrancy guard patterns) was provided to the LLMs in all relevant experiments to ensure consistency. Figure 6.1 illustrates the complete methodological workflow.

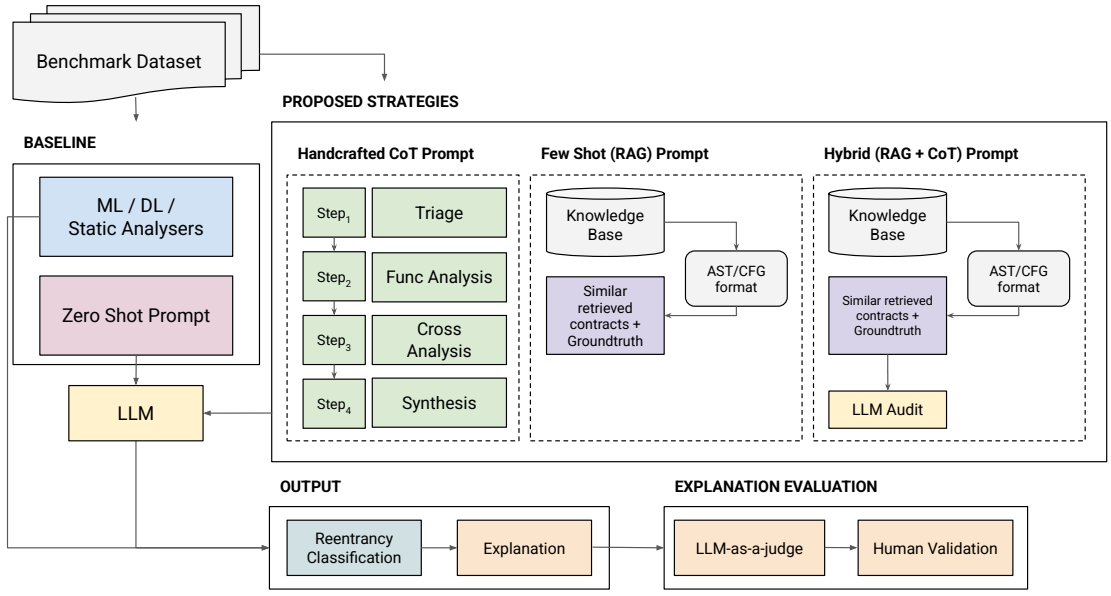



Figure 6.1: Methodology workflow.

### 6.5.1 Reentrancy Detection Principles

To ensure a consistent and rigorous analytical standard across all experimental arms—from human expert validation to the LLM-based strategies—I established a formal set of principles for reentrancy detection. This framework moves beyond naive pattern matching to incorporate a nuanced understanding of smart contract execution logic, mitigation techniques, and plausible exploitability.

At its core, this framework is built upon a strict definition of the CEI pattern. A contract is considered to adhere to this pattern if all state modifications are unconditionally completed *before* any external call within a given logical operation. For this purpose, an *Effect* is strictly defined as a contract state modification via the assignment operator. Operations such as event emissions or `require`/`assert` statements are not considered Effects.

Conversely, an *Interaction* is narrowly defined as an external call that can shift control flow to a potentially malicious contract. This includes low-level primitives like `.call` and `.delegatecall`, as well as any method invocation on an external contract or interface type. Primitives that do not transfer execution control in a reentrant manner, such as

`.staticcall`, or that have built-in gas limitations that prevent reentrancy, like native `.send` and `.transfer`, are explicitly excluded from this definition.

The presence of a potential CEI violation is a necessary, but not sufficient, condition for vulnerability classification. The proposed framework requires identifying a *plausible exploit path* where reentrancy leads to a tangible, harmful outcome, which may include not only the theft of assets but also critical state inconsistencies that break contract logic.

Finally, the framework accounts for common mitigation strategies. A function protected by a correctly implemented and applied reentrancy guard (e.g., a standard `nonReentrant` modifier or a custom mutex) is generally considered safe from re-entering itself. However, in cases of *cross-function reentrancy*—where an external call in function `A` allows re-entry into a different function `B` that shares state with `A`—the protective mechanism must correctly guard all relevant functions in the potential execution path to be considered effective.

### 6.5.2 Benchmark Construction and Validation

The cornerstone of this investigation is a benchmark dataset constructed to address critical deficiencies in extant resources. An initial corpus was aggregated from three established sources [39, 32, 66] and subsequently underwent deduplication and compilation filtering. The principal contribution resides in a manual verification phase conducted by three domain experts, following a predefined rubric based on a formal definition of reentrancy. This audit revealed profound inaccuracies in prior labels: 28 contracts previously designated as reentrant were confirmed safe, whereas 5 labeled safe contained vulnerabilities. To quantify the reliability of this process, I measured the inter-rater reliability before a final consensus discussion, achieving a Fleiss' Kappa of 0.89, indicating substantial agreement. This exacting process yielded a final benchmark of 436 contracts (122 reentrant, 314 safe), providing a high-fidelity ground truth.

To contextualize the performance of the proposed approach, I compare it against a diverse set of strong baselines spanning traditional program analysis, classical and neural ML, and recent zero-shot language models. To benchmark against established non-learning techniques, I evaluated a portfolio of three prominent analyzers: *Slither*, *Mythril*, and *Confuzzius*. These were deliberately selected to represent the field's dominant paradigms: static analysis, symbolic execution, and fuzzing, respectively. The tools were run using their default configurations via the SmartBugs framework to establish a fair and reproducible benchmark of their standard, out-of-the-box performance.

I also investigated seven traditional ML models (Gradient Boosting, Gaussian Naive Bayes, K-Nearest Neighbors, Logistic Regression, Random Forest, Support Vector Machine, and Extreme Gradient Boosting) and three DL architectures (Feed Forward Neural Network, Bidirectional Long Short-Term Memory, and CodeBERT). To ensure competitive performance, all traditional models were hyperparameter-optimized using a grid search.

For LLMs, I evaluated the untuned capabilities of six models: GPT-4o, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, o3-mini, and o4-mini. I deliberately exclude fine-tuning to isolate the effects of in-context learning, reserving a direct comparison against it for future work. To enhance LLM performance beyond zero-shot capabilities, I explore three advanced prompting strategies designed to inject structural priors, expert reasoning patterns, or both. These approaches aim to systematically guide the model toward accurate and interpretable vulnerability assessments.

**Structurally-Aware RAG.** This strategy grounds the LLM with in-context examples. For a target contract, the proposed pipeline retrieves the top-k structurally analogous contracts ($k = 3$, determined empirically in Section 6.6.1) based on the graph similarity of their ASTs and CFGs. The source code of these examples is then provided to the LLM as context for its analysis.

**Expert-Crafted CoT.** This strategy implements the four-step audit process not as a single monolithic prompt, but as a programmatic chain of four sequential LLM calls, where the output of one step is consumed as the input for the next. The sequence enforces a formal decomposition of the problem: (1) a *Triage* step identifies all functions with external calls; (2) a localized *Function Analysis* step assesses each function individually; (3) a *Cross Function Analysis* step analyzes the interaction between these functions; and (4) a final *Synthesis* step aggregates all intermediate findings into a conclusive verdict.

**Hybrid (CoT + RAG) Strategy.** This strategy integrates evidence with reasoning through a two-stage context enrichment process. First, in a preliminary step, a separate LLM instance generates a detailed security audit for each of the $k = 3$ retrieved examples and their corresponding labels, explaining *why* they are safe or vulnerable according to a standardized template. Second, the primary LLM is tasked with analyzing the target contract, but it is provided with these pre-computed, structured analyses as its context, rather than raw source code. This approach provides a set of worked examples that demonstrate how to apply the reasoning principles to concrete cases before the model begins its analysis.

To rigorously evaluate model performance, I designed a two-part protocol that considers both *predictive accuracy* and the *quality of generated explanations*. This ensures a comprehensive assessment—capturing not only what the models predict, but also how and why they arrive at their conclusions. Predictive performance was measured using standard macro-averaged metrics—precision, recall, and F1 score — calculated on a 3-fold cross-validation setup to ensure robustness across samples. To assess the quality of explanations, I adopted a two-stage process. First, an automated evaluation used an *LLM-as-a-Judge* (specifically, `o4-mini`) to evaluate the explanations against a structured rubric. This rubric was aligned with the same formal criteria used during task design. It evaluated each explanation across three dimensions: correctness (factual accuracy), informativeness (depth and helpfulness), and pertinence (conciseness and relevance). In the second stage, three security professionals evaluated a randomly selected sample of 88 explanations from one of the cross-validation test splits. The human experts were required to ground their evaluation on the same rubric employed by the LLM-as-a-Judge. I used the two-sided Wilcoxon signed-rank test ($p < 0.05$) to assess statistical significance in the ordinal ratings.

The meticulous curation of the proposed benchmark serves not only to validate experimental results but also addresses a broader reproducibility crisis in ML-based vulnerability detection. By documenting labeling ambiguities, conducting multi-expert validation, and providing a high-fidelity dataset, this work sets a precedent for rigor in smart contract security research. Such a resource enables consistent evaluation, fosters method comparability, and supports the development of more robust, generalizable detection techniques.

### 6.5.3 Implementation and Reproducibility

Implementations relied on standard libraries: traditional models used `scikit-learn (v1.6.0)`; DL models used `PyTorch (v2.4.1)`. The BiLSTM/FFNN models were trained for 50 epochs (batch size 32, AdamW, LR $1 \times 10^{-4}$), while CodeBERT was fine-tuned for 5 epochs (batch size 16, LR $2 \times 10^{-5}$) from the `microsoft/codebert-base`

checkpoint.

The RAG pipeline used `Slither (v0.9.0)` for AST/CFG generation and `GraKeL (v0.1.8)` with the Weisfeiler-Lehman kernel (3 iterations) to measure graph similarity. All LLM interactions were conducted at a temperature of 0. To ensure full transparency and facilitate replication, all source code, data, and the precise, structured prompt templates used to implement the described strategies are publicly available in the code repository associated with this work.

## 6.6 Results

I present the evaluation results of this work along two dimensions: classification accuracy and explanation quality. Accuracy compares the predictive performance of traditional, neural, static analysis, and generative models, while explanation quality—measured via automated metrics and a user study—assesses correctness, informativeness, and relevance.

### 6.6.1 Example Retrieval Optimization

The effectiveness of the RAG approach critically depends on retrieving an appropriate number of relevant examples, controlled by the parameter $k$. Selecting an optimal $k$ requires balancing the benefits of contextual information against the risk of introducing noise or conflicting examples. To determine the best value, I conducted a sensitivity analysis by varying $k$ from 1 to 7 and measuring its effect on classification accuracy and preliminary explanation quality, using GPT-4.1-nano on a single test split.

These experiments were conducted using the RAG setup across different input data representations derived from AST, CFG, and their combination (i.e., a weighted average of the individual similarity scores). The findings of this work, summarized graphically in Figure 6.2, indicate a consistent trend: model performance generally improves as $k$ increases from 1 to 3. However, for $k > 3$, I observed a tendency for F1 performance to degrade. This suggests that while retrieving a small number of highly relevant examples ($k \leq 3$) enhances the model's context, incorporating more examples ($k > 3$) increasingly introduces less pertinent information, potentially confusing the model or diluting the signal from the most relevant retrieved documents. Based on this empirical analysis, I selected $k = 3$ as the optimal number of retrieved examples for all subsequent RAG experiments reported in this study. This value represents the best-observed trade-off between providing sufficient context and minimizing noise.

To enhance the Structurally-Aware RAG pipeline, I conducted preliminary experiments to determine which structural representation most effectively captures contract similarity. This work's central hypothesis is that the choice of structural encoding directly influences the relevance of retrieved examples, which in turn affects the LLM's downstream analysis. I evaluated three representations: the AST, the CFG, and a combined AST+CFG approach (using averaged similarity scores). For clarity, I report results for GPT-4.1, GPT-4.1-mini, and GPT-4.1-nano in Table 6.1, representing large, medium, and small model sizes, respectively.

Across all model scales, I observed a consistent trend: retrieval based on CFG similarity outperforms both AST and combined representations. This advantage is particularly pronounced for GPT-4.1-nano and GPT-4.1-mini, and remains stable even for the larger

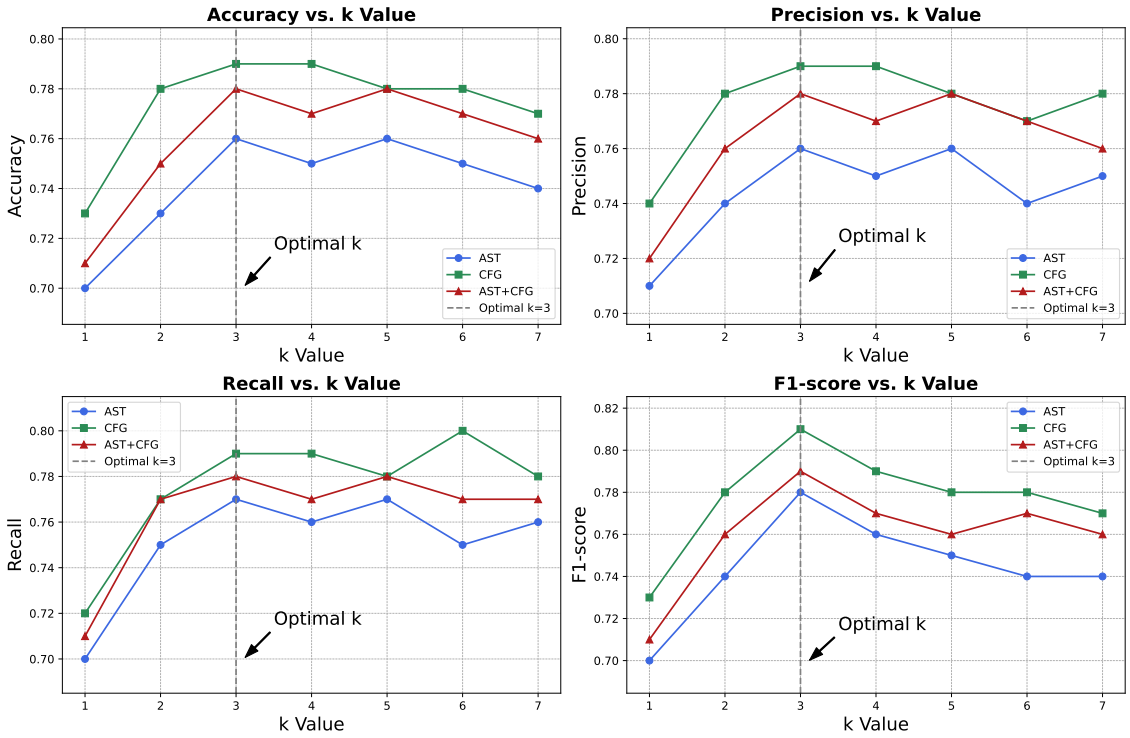



Figure 6.2: Impact of the number of retrieved examples ($k$) on model performance across different data representations.

GPT-4.1 model, where CFG-based retrieval yields the highest F1 score with the lowest standard deviation.

This result highlights an important insight: while vulnerability detection is ultimately performed by the LLM analyzing the full source code, the retrieval stage plays a crucial role in shaping the quality of that analysis. In the case of reentrancy, which depends on the execution order of operations—such as an external call preceding a state update—the CFG provides a better signal for identifying structurally similar contracts. Although this work's CFGs are not semantically annotated, the node labels generated by Slither (e.g., indicating low-level calls) are preserved in the graph and leveraged during similarity computation via the Weisfeiler-Lehman kernel. As a result, the CFG captures both execution flow and lightweight semantic cues that guide retrieval more effectively than syntax-based approaches.

Interestingly, considering both AST and CFG similarities did not improve performance over CFG alone. I hypothesize that this may be due to AST introducing redundant or less relevant information, thereby diluting the more task-relevant control-flow signals captured by the CFG. Additionally, without a more sophisticated fusion method, the averaged representation may blur the distinct structural signals, ultimately reducing retrieval precision.

Given these findings, I selected the CFG as the default structural representation for all subsequent experiments involving RAG and Hybrid retrieval strategies.

These results highlight an important principle: the structural representation of code substantially shapes the LLM's reasoning pathway. By aligning retrieved examples with control-flow-sensitive patterns, the model more accurately internalizes causal relationships critical for reentrancy detection. This observation supports the thesis hypothesis that structural priors are not merely auxiliary but central to trustworthy LLM-guided auditing,

particularly in domains where order-of-execution errors have severe consequences.

Table 6.1: RAG strategies using different structural representations. Values are reported as Mean (Standard Deviation). The best result for each model is in bold.

| Model | Representation | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| GPT-4.1 | AST + CFG | 0.91 (0.03) | 0.92 (0.03) | 0.91 (0.03) | 0.91 (0.03) |
| | AST | 0.92 (0.03) | 0.92 (0.03) | 0.92 (0.03) | 0.92 (0.03) |
| | CFG | **0.92(0.02)** | **0.92(0.01)** | **0.92(0.02)** | **0.92(0.01)** |
| GPT-4.1-mini | AST + CFG | 0.85 (0.00) | 0.86 (0.02) | 0.85 (0.00) | 0.84 (0.00) |
| | AST | 0.85 (0.01) | 0.86 (0.01) | 0.85 (0.01) | 0.84 (0.02) |
| | CFG | **0.88(0.01)** | **0.88(0.01)** | **0.88(0.01)** | **0.88(0.01)** |
| GPT-4.1-nano | AST + CFG | 0.74 (0.01) | 0.75 (0.04) | 0.74 (0.01) | 0.66 (0.01) |
| | AST | 0.73 (0.02) | 0.73 (0.07) | 0.73 (0.02) | 0.66 (0.02) |
| | CFG | **0.81(0.01)** | **0.82(0.02)** | **0.81(0.01)** | **0.79(0.01)** |

### 6.6.2 Models Accuracy

This work's empirical evaluation reveals a clear hierarchy of efficacy among competing reentrancy-detection methods. The findings, summarized in Table 6.2, not only quantify the performance of different approaches but also yield critical insights into the trade-offs between model architecture, prompting strategy, and computational cost.

The central finding of this study is that LLMs, when properly guided, establishes a new SotA. A crucial distinction emerges between the general-purpose `GPT` series and the reasoning-optimized o series. While the larger GPT-4o is a powerful generalist, the o3-mini model, particularly when augmented with the proposed Structurally-Aware RAG strategy, achieved the highest overall F1 Score.

Such a performance divergence stems from their different design philosophies. `GPT` models are optimized for broad applicability, whereas o models are explicitly trained to excel at multi-step logical problems. This inherent specialization gives `o3-mini` a significant advantage, allowing it to achieve an exceptional baseline score even without complex prompting.

This distinction also explains the nuanced impact of this work's prompting strategies. The CoT framework, designed to impose a logical structure, provides a clear benefit to some generalist `GPT` models but shows diminishing returns for the o series. This suggests that the o series, having been trained to generate its own optimized reasoning paths, can experience "procedural interference" from an externally enforced workflow.

Conversely, while RAG provided the highest performance ceiling, its universal effectiveness is not absolute. The primary performance bottleneck for top-tier models was a lack of specific domain knowledge, which RAG directly addresses. However, for smaller models, the core limitation may be reasoning capacity itself, a factor that RAG cannot fully remediate. The frequent failure of the hybrid RAG+CoT strategy to outperform RAG alone further suggests a phenomenon of *constraint-induced sub-optimality*, where a rigid procedure limits a powerful model's ability to synthesize the rich information flexibly embedded by RAG-provided exemplars.

Table 6.2: Overall performance comparison across all model families and strategies. Learning-based model results are reported as Mean (Standard Deviation), best per strategy in *italic* and overall best in **bold**.

| Approach | Strategy | Accuracy | Precision | Recall | F1 Score |
|---|---|---|---|---|---|
| *Large Language Models* | | | | | |
| o3-mini | Baseline | 0.96 (0.01) | 0.96 (0.01) | 0.96 (0.01) | 0.96 (0.01) |
| | Baseline + CoT | 0.94 (0.01) | 0.95 (0.01) | 0.94 (0.01) | 0.95 (0.01) |
| | RAG | **0.97(0.01)** | **0.97(0.01)** | **0.97(0.01)** | **0.97(0.01)** |
| | RAG + CoT | 0.94 (0.01) | 0.95 (0.01) | 0.94 (0.01) | 0.94 (0.01) |
| o4-mini | Baseline | 0.93 (0.02) | 0.94 (0.01) | 0.93 (0.02) | 0.93 (0.02) |
| | Baseline + CoT | *0.95 (0.01)* | *0.95 (0.01)* | *0.95 (0.01)* | *0.95 (0.01)* |
| | RAG | 0.94 (0.02) | 0.94 (0.02) | 0.94 (0.02) | 0.94 (0.02) |
| | RAG + CoT | 0.94 (0.02) | 0.95 (0.02) | 0.94 (0.02) | 0.94 (0.02) |
| GPT-4.1 | Baseline | 0.87 (0.02) | 0.90 (0.01) | 0.87 (0.02) | 0.88 (0.02) |
| | Baseline + CoT | 0.87 (0.03) | 0.90 (0.02) | 0.87 (0.03) | 0.87 (0.02) |
| | RAG | *0.92 (0.02)* | *0.92 (0.01)* | *0.92 (0.02)* | *0.92 (0.01)* |
| | RAG + CoT | *0.92 (0.01)* | *0.92 (0.00)* | *0.92 (0.01)* | *0.92 (0.01)* |
| GPT-4.1-mini | Baseline | 0.86 (0.01) | 0.88 (0.01) | 0.86 (0.01) | 0.85 (0.02) |
| | Baseline + CoT | *0.90 (0.01)* | *0.91 (0.02)* | *0.90 (0.01)* | *0.90 (0.01)* |
| | RAG | 0.88 (0.01) | 0.88 (0.01) | 0.88 (0.01) | 0.88 (0.01) |
| | RAG + CoT | 0.88 (0.01) | 0.88 (0.01) | 0.88 (0.01) | 0.88 (0.01) |
| GPT-4.1-nano | Baseline | 0.72 (0.00) | 0.61 (0.13) | 0.72 (0.00) | 0.61 (0.01) |
| | Baseline + CoT | 0.71 (0.02) | 0.74 (0.01) | 0.71 (0.02) | 0.72 (0.02) |
| | RAG | *0.81 (0.01)* | *0.82 (0.02)* | *0.81 (0.01)* | *0.79 (0.01)* |
| | RAG + CoT | 0.80 (0.01) | 0.80 (0.01) | 0.80 (0.01) | 0.77 (0.02) |
| GPT-4o | Baseline | 0.85 (0.02) | 0.89 (0.02) | 0.85 (0.02) | 0.86 (0.02) |
| | Baseline + CoT | 0.86 (0.03) | 0.90 (0.02) | 0.86 (0.03) | 0.86 (0.03) |
| | RAG | *0.92 (0.01)* | 0.92 (0.01) | *0.92 (0.01)* | 0.92 (0.01) |
| | RAG + CoT | *0.92 (0.01)* | *0.93 (0.01)* | *0.92 (0.01)* | *0.93 (0.01)* |
| *Traditional and DL Baselines* | | | | | |
| Traditional ML | Gradient Boosting | 0.90 (0.04) | 0.87 (0.04) | 0.76 (0.11) | 0.81 (0.05) |
| | Gaussian NB | 0.82 (0.05) | 0.70 (0.06) | 0.65 (0.12) | 0.67 (0.07) |
| | KNN | 0.88 (0.01) | 0.85 (0.13) | 0.74 (0.13) | 0.78 (0.01) |
| | Logistic Regression | 0.80 (0.06) | 0.93 (0.12) | 0.33 (0.12) | 0.47 (0.12) |
| | Random Forest | 0.90 (0.04) | 0.90 (0.07) | 0.73 (0.13) | 0.80 (0.07) |
| | SVM | 0.84 (0.04) | **0.93(0.12)** | 0.51 (0.16) | 0.64 (0.09) |
| | XGBoost | *0.91 (0.02)* | 0.88 (0.08) | *0.79 (0.10)* | *0.83 (0.02)* |
| Deep Learning | CodeBERT | 0.90 (0.13) | 0.82 (0.24) | **0.96(0.03)** | **0.87(0.14)** |
| | LSTM | 0.86 (0.12) | 0.76 (0.17) | **0.96(0.03)** | 0.83 (0.11) |
| | FFNN | **0.96(0.01)** | *0.85 (0.03)* | 0.89 (0.03) | 0.86 (0.02) |
| *Static Analysis Tool Baselines* | | | | | |
| Static Analysis Tools | Confuzzius | **0.88** | **0.94** | 0.63 | **0.75** |
| | Mythril | 0.75 | 0.61 | 0.31 | 0.41 |
| | Slither | 0.86 | 0.82 | **0.64** | 0.72 |

#### 6.6.2.1 Discussion

The inclusion of the smaller GPT-4.1-mini and GPT-4.1-nano models illuminates a critical dimension for practical application: the trade-off between performance and resource efficiency. These faster, cheaper models provide valuable insights into the minimum viable capabilities for this task.

The **GPT-4.1-mini** model emerges as a highly compelling option. With a simple CoT strategy, it achieved an F1 score that outperformed the baseline results of the much larger GPT-4.1 and GPT-4o models. This finding is significant, as it suggests that for moderately sized models, a structured reasoning framework (CoT) can be more effective than providing raw knowledge (RAG), likely because it scaffolds the model's more limited intrinsic reasoning capabilities. For organizations where the cost and latency of flagship models are prohibitive, GPT-4.1-mini with CoT represents a "sweet spot", offering robust performance that far exceeds traditional baselines at a fraction of the computational budget.

The **GPT-4.1-nano** model defines the lower bound of effectiveness. With a top F1-score using RAG, its performance is substantially weaker than all other LLMs. This demonstrates that there is a "performance floor"—a minimum threshold of model scale and complexity required to move beyond surface-level pattern matching to the deeper semantic analysis necessary for this task. Even the best prompting strategy cannot fully compensate for a fundamental lack of reasoning power. Tellingly, the nano model's performance is only marginally better than the best static analysis baseline in this work (i.e., Confuzzius), highlighting that at this small scale, the LLM's advantage begins to erode.

As a final remark, I delve now into a comparative analysis against this work's baseline models, which elucidates the specific failure modes of previous-generation techniques and underscores the nature of the LLM's advantage.

The DL baselines, such as `CodeBERT`, exhibited a characteristic high-recall, low-precision profile, achieving a recall of $0.96$ but a much lower F1-score. This pattern suggests that such models are adept at learning the *syntactic correlates* of vulnerabilities—for example, the presence of an external call within a function—but fail to grasp the deeper, temporal semantics required to distinguish a benign call from a reentrant one (i.e., whether the call occurs *before* a state update). In contrast, the high F1-scores of appropriately prompted LLMs imply they are successfully capturing this essential logic.

The static analysis tools, foundational to current security practices, were significantly outperformed by nearly all learning-based models. Their performance, characterized by low F1-scores and particularly poor recall, highlights the inherent brittleness of heuristic-based detection. These tools rely on predefined patterns that generalize poorly. This "semantic gap" between rigid heuristics and complex reality is precisely what modern LLMs, even the cost-effective GPT-4.1-mini, are equipped to bridge.

### 6.6.3 Explainability

Beyond predictive accuracy, the ultimate value of an AI security tool lies in the quality of its explanations. A correct classification is of limited utility in a high-stakes security context if its underlying reasoning is flawed, opaque, or untrustworthy. To rigorously assess this critical dimension, we employed a two-pronged evaluation strategy: combining nuanced, definitive ratings from human security experts with scalable, fine-grained assessments from an LLM-as-a-Judge framework. This dual approach provides a compre-

hensive picture of the practical utility of the generated explanations.

Our human expert evaluation (Table 6.3) immediately confirmed that the method of generating an explanation profoundly shapes its value. The Structurally-Aware RAG strategy produced explanations judged as significantly more useful than the baseline zero-shot approach, especially in Informativeness (4.72 vs. 4.14). This quantitative gap signifies a fundamental qualitative shift in function. A baseline explanation delivers a verdict—a claim a developer must then independently verify. In contrast, an RAG-generated explanation delivers a verdict along with verifiable evidence in the form of an analogous, real-world smart contract. This transforms the AI's output from an unsubstantiated claim into a pedagogical tool. It illuminates the why behind a vulnerability, builds developer intuition, and directly accelerates the remediation workflow.

This evidence-grounding is also paramount for establishing user trust. The superior Correctness score for RAG demonstrates that grounding the LLM in concrete examples measurably mitigates the risk of factual hallucination, bolstering developer confidence in the analysis. Interestingly, the baseline's marginally higher Pertinence score (4.37 vs. 4.31) reveals a subtle but important user preference for conciseness. This highlights a tension between detail and brevity, a theme that our automated evaluation explores in greater detail.

Table 6.3: Comparison of Qualitative Explanation Metrics: BASELINE vs. RAG o3-mini. Values are reported as Mean (Standard Deviation) on a 1-5 scale. The best performing strategy for each metric is highlighted in **bold**.

| Metric | Baseline | RAG |
|---|---|---|
| Correctness | 4.38 (0.94) | **4.43(0.84)** |
| Informativeness | 4.14 (0.70) | **4.72(0.57)** |
| Pertinence | **4.37(0.87)** | 4.31 (0.84) |

#### 6.6.3.1 Discussion

The LLM-as-a-Judge assessment (Table 6.4) allows us to dissect how different prompting strategies interact with diverse model architectures at scale. The results crystallize several key principles for designing explainable AI systems.

First, evidence-grounding is a universally dominant principle. Across all models, from the high-capacity o3-mini to the compact GPT-4.1-nano, the Structurally-Aware RAG strategy consistently yielded the best-balanced explanations. Its power to uplift performance is particularly stark for smaller models; for example, RAG elevated GPT-4.1-nano's Correctness score from a mediocre 3.53 to a robust 4.42. This proves that providing verifiable, contextual evidence is the most reliable path to generating trustworthy explanations, effectively establishing a high-quality floor regardless of model size.

Second, the utility of Chain-of-Thought (CoT) is highly model-dependent. For the reasoning-optimized o models, the explicit step-by-step logic of CoT offered minimal benefit and often reduced Pertinence. This supports our "procedural interference" hypothesis: these specialized models already possess streamlined internal reasoning pathways, and the rigid scaffolding of CoT can be redundant or even counterproductive. Conversely, the generalist GPT models showed marked improvements in Informativeness with CoT, suggesting that, for less specialized architectures, this explicit structure is essential for producing coherent, logical output.

Finally, the analysis confirms a foundational trade-off between Informativeness and Pertinence. The hybrid RAG + CoT strategy consistently produced the most detailed and exhaustive explanations across all models, earning the highest Informativeness scores. However, this same verbosity consistently yielded some of the lowest Pertinence scores. This is not an anomaly but a systemic feature of the methodology, presenting a critical design tension: maximizing explanatory detail can directly undermine its conciseness and, therefore, its immediate utility.

In synthesis, our dual evaluation converges on a central, actionable insight. The most effective AI security tools are not those that attempt to mimic the entirety of a human's internal thought process (CoT), but those that emulate an expert's practical use of evidence (RAG). By grounding its analysis in concrete, verifiable examples, the Structurally-Aware RAG strategy produces explanations that are correct, insightful, and efficiently actionable. This fosters a collaborative partnership between the developer and the AI, transforming the tool from a black-box oracle into a trusted co-pilot for secure software development.

## 6.7 Limitations

Several limitations define the scope of the findings in this work. This study focuses exclusively on reentrancy in Solidity; thus, the generalizability of its methods to other vulnerability classes or blockchain ecosystems remains an open question. The reliance on proprietary, black-box LLMs presents challenges for long-term reproducibility and introduces practical constraints, such as cost and latency, for real-time applications. Finally, the methodology of this work is limited to static source code analysis. It does not account for dynamic on-chain states or complex multi-contract interactions, which can be sources of emergent exploits.

## 6.8 Conclusions

This work demonstrates that LLMs, when coupled with the proposed Structurally-Aware RAG strategy, achieve a new SotA in detecting reentrancy vulnerabilities in smart contracts, surpassing traditional ML, DL, and static analysis baselines. The results reveal that, for complex program analysis, grounding generative reasoning in factual and structural precedents is more effective than enforcing rigid, self-contained reasoning chains. This evidence-centric strategy not only enhances predictive accuracy but also yields intelligible, authentic explanations that are grounded in the underlying code logic and pragmatically actionable. Beyond performance, these findings articulate a broader insight. They suggest that LLMs, when properly contextualized and structurally grounded, can serve as systems capable of co-constructing understanding rather than merely automating detection. The Structurally-Aware RAG approach exemplifies how explainability can be embedded as a design principle rather than appended as a post-hoc justification, thereby operationalizing the tripartite framework of *faithfulness*, *intelligibility*, and *alignment*. Within this framework, retrieval grounding enhances faithfulness by linking outputs to verifiable causal structures, while structured reasoning improves intelligibility by producing explanations that align with human cognitive schemas. Additionally, prompting strategies ensure alignment by shaping the communicative interface between the model and the expert.

Viewed through this theoretical lens, the empirical results extend beyond blockchain security. They demonstrate a pathway toward reconciling the generative capacities of

Table 6.4: Per-Model LLM-as-a-Judge Evaluation of Generated Explanations. Values are reported as Mean (Standard Deviation) on a 1-5 scale. The best-performing strategy for each model and metric is highlighted in *italic*, with the overall best strategy in **bold**.

| Model | Strategy | Correctness | Informativeness | Pertinence |
|---|---|---|---|---|
| o3-mini | Baseline | 4.61 (0.75) | 4.03 (0.90) | 4.71 (0.60) |
| | Baseline + CoT | 4.72 (0.65) | 4.55 (0.70) | 4.45 (0.80) |
| | RAG | **4.91(0.40)** | 4.83(0.50) | **4.85(0.45)** |
| | RAG + CoT | 4.84 (0.45) | **4.93(0.40)** | 4.22 (0.95) |
| o4-mini | Baseline | 4.43 (0.80) | 3.81 (1.05) | 4.65 (0.65) |
| | Baseline + CoT | 4.65 (0.70) | 4.41 (0.80) | 4.31 (0.85) |
| | RAG | *4.83 (0.50)* | 4.71 (0.60) | *4.77 (0.55)* |
| | RAG + CoT | 4.76 (0.55) | *4.81 (0.50)* | 4.15 (1.00) |
| GPT-4.1 | Baseline | 3.82 (1.10) | 3.05 (1.20) | 4.21 (0.85) |
| | Baseline + CoT | 4.25 (0.95) | 4.08 (0.90) | 4.03 (0.90) |
| | RAG | *4.71 (0.65)* | 4.54 (0.75) | *4.63 (0.70)* |
| | RAG + CoT | 4.64 (0.70) | *4.66 (0.65)* | 3.88 (1.15) |
| GPT-4.1-mini | Baseline | 3.91 (1.05) | 3.25 (1.15) | 4.25 (0.80) |
| | Baseline + CoT | 4.33 (0.90) | 4.15 (0.95) | 4.08 (0.85) |
| | RAG | *4.68 (0.68)* | 4.51 (0.78) | *4.66 (0.65)* |
| | RAG + CoT | 4.55 (0.72) | *4.63 (0.68)* | 3.95 (1.10) |
| GPT-4.1-nano | Baseline | 3.53 (1.20) | 2.51 (1.30) | 4.05 (0.95) |
| | Baseline + CoT | 3.88 (1.10) | 3.55 (1.10) | 3.81 (1.00) |
| | RAG | *4.42 (0.80)* | 4.23 (0.90) | *4.45 (0.80)* |
| | RAG + CoT | 4.31 (0.85) | *4.35 (0.80)* | 3.64 (1.25) |
| GPT-4o | Baseline | 3.75 (1.15) | 2.89 (1.25) | 4.13 (0.90) |
| | Baseline + CoT | 4.11 (1.00) | 3.84 (1.00) | 3.91 (0.95) |
| | RAG | *4.65 (0.70)* | 4.42 (0.80) | *4.55 (0.75)* |
| | RAG + CoT | 4.58 (0.75) | *4.59 (0.70)* | 3.71 (1.20) |

modern LLMs with the epistemic demands of explainability. In doing so, they respond to the central question animating this thesis: not whether machines can think, but whether humans can understand how machines think.

Several research avenues naturally follow. A key direction is to assess the generalizability of this framework across diverse vulnerability classes and programming languages, testing its robustness and scope. Methodologically, integrating richer program representations—such as Program Dependence Graphs or syntax-aware embeddings—could further enhance the faithfulness of retrieval and reasoning. Extending the current one-shot pipeline into an interactive, human-in-the-loop auditing system would enable security experts to iteratively query, refine, and validate AI-generated insights, thereby materializing the thesis's vision of *co-evolutionary explainability*. Finally, optimizing latency and computational cost through model distillation or retrieval-efficient architectures will be essential for real-time deployment in continuous integration environments, ensuring that high-fidelity AI auditing remains both scalable and accountable.

In sum, this chapter demonstrates that combining LLMs with structured, evidence-grounded retrieval not only advances the state of smart contract analysis but also concretizes the thesis's broader argument: that explainability, when treated as an epistemic design principle, enables a new synthesis between machine reasoning and human understanding. In the generative era, this synthesis is not peripheral to intelligence—it *is* intelligence. The next chapter will build on these foundations to explore the role of explainability in model design.

# 7

# Assessing the Industrial Stakes of Explainable Artificial Intelligence

***This chapter is based on:*** *M. Rizzo et al. 'Leveraging periodicity for tabular deep learning'. In:* Electronics *14.6 (Mar. 2025), p. 1165. ISSN: 2079-9292. DOI: 10.3390/electronics14061165. URL: https://doi.org/10.3390/electronics14061165*

As AI systems move from research into production, the question of how to explain their behavior becomes inseparable from their governance. In high-stakes industrial sectors—from finance and energy to healthcare and manufacturing—AI models make decisions with significant material and ethical consequences. Here, XAI is not merely a diagnostic tool but a structural requirement, enabling organizations to trust, audit, and control algorithmic reasoning at scale.

As shown in previous chapters, the epistemic function of explanation is context-dependent. In medicine, explanations are primarily *communicative* acts that bridge the gap between an AI model's representations and a clinician's diagnostic heuristics. In the adversarial context of smart contract security, however, explanations become tools for *knowledge construction*. They help analysts surface vulnerabilities and reason about system behavior, operating as a generative mechanism for creating new interpretive frameworks. These cases reveal a clear trajectory for the role of XAI: from *communication* to *knowledge construction* and ultimately to *structural design*. The industrial context extends this evolution, positioning explainability as a core principle for engineering large-scale systems where accountability, safety, and performance are intertwined.

## 7.1 Explainability as an Industrial Design Principle

This thesis argues that explanations must be both faithful to a model's internal logic and aligned with the epistemic structures of the application domain. This alignment transforms explainability from a post hoc justification into a mechanism for ensuring continuity between automated inference and expert reasoning.

In industrial practice, this principle operates on three interdependent levels. *Strategically*, explainability supports governance and regulatory compliance under frameworks like the EU AI Act. *Operationally*, it serves as an interpretive interface between model outputs and human experts, who must diagnose, validate, and refine algorithmic behavior. *Technically*, it informs the model architecture itself, motivating designs that are structurally grounded in the domain's natural regularities.

The ultimate challenge, therefore, is not simply to explain pre-trained models but to engineer models that are explainable by design. This goal is particularly acute in industry, where the dominant data modality—tabular data—lacks the intuitive spatial or semantic regularities of images or text, making interpretation difficult.

## 7.2 Tabular Data: The Industrial Frontier of XAI

Most industrial data pipelines are built on tabular data—structured records of measurements, transactions, or system states. Despite its ubiquity, this data modality has received comparatively little attention from the XAI community. Without the clear spatial or semantic features found in other data types, understanding tabular models requires uncovering the latent statistical and temporal structures that organize the dataset. The core challenge is to illuminate the generative structure of the data: the recurrent dependencies, systemic couplings, and periodic fluctuations that drive observed behavior.

Among these latent structures, periodicity is particularly central and explainable in industrial systems. Energy grids, manufacturing lines, and financial markets all exhibit cyclical dynamics that encode system stability and disruption. Capturing these patterns provides a direct path to in-model explainability, in which the model's architecture encodes mechanisms that correspond to established domain knowledge, rather than relying on external attribution methods. This insight motivates the core approach of this chapter. By integrating periodic representations directly into their architectures, models can make their reasoning more naturally intelligible.

## 7.3 Leveraging Periodicity for Explainable Predictions in Tabular Scenarios

The workhorse of industrial AI is *tabular data* — structured datasets composed of rows and columns, encompassing domains as varied as finance, healthcare, manufacturing, marketing, and the social sciences [169]. These datasets encode the logic of real-world processes, such as patient trajectories in medical records, consumption cycles in energy markets, or transaction histories in finance. Their ubiquity makes them a privileged site for treating explainability as a question of how model reasoning can be made congruent with domain understanding.

Despite remarkable advances in DL for unstructured data, its adoption in tabular domains has lagged. Traditional ML methods, such as Gradient Boosting Machines and Random Forests, often outperform deep architectures on tabular datasets [64]. This discrepancy arises from the absence of the spatial or sequential regularities that neural architectures like CNNs or RNNs exploit [89], as well as from the heterogeneous and hierarchically entangled nature of tabular features [210]. Consequently, deep models for tabular data tend to learn representations that are not only data-hungry but also *opaque*—difficult to interpret and validate in operational contexts.

Several recent architectures—such as TabNet [8], NODE [142], DeepGBM [91], and FT-Transformer [58, 64]—have sought to close this gap through attention mechanisms and differentiable decision trees. Yet, while these models improve accuracy, they often reproduce the fundamental opacity of end-to-end neural reasoning. Explanations, when attempted post hoc, remain descriptive rather than structural: they tell us *which features* influenced a prediction, not *why* the model relied on them or how this reasoning relates to the generative structure of the data.

From the perspective of this thesis, this limitation is not merely technical but epistemological. It exemplifies the broader crisis of faithfulness in XAI: explanations detached from model causality or domain logic risk becoming rhetorical artifacts. To overcome this, I seek forms of explainability that are *constitutive* rather than corrective—built into the model's representational fabric. This motivates exploring *periodicity* as an explainable structural prior for tabular data in this work.

Many real-world datasets, even those without explicit temporal attributes, exhibit *latent periodicity*—regular, repeating patterns that reflect underlying rhythms of human or physical processes [40]. Examples include daily consumption cycles, weekly production shifts, seasonal market fluctuations, or circadian biological variations. Traditional deep tabular models often capture these patterns implicitly but cannot explicitly articulate them. As a result, their decisions may be accurate yet inaccessible: the model "knows" that a sales peak is seasonal but cannot explain that reasoning in a form intelligible to human experts.

The proposed approach reframes periodicity not only as a predictive cue but as a vehicle for explainability. By embedding cyclical structures into the model's representational space, I make the statistical regularities of the data coextensive with human-understandable temporal logics. In other words, the model's reasoning becomes structurally aligned with the epistemic structures of the domain—a practical manifestation of the *faithfulness–intelligibility–alignment* triad articulated earlier in this thesis.

I therefore introduce neural architectures that disentangle and explicitly model both periodic and non-periodic patterns in tabular data. Specialized encoding techniques are used to capture cyclical dependencies via Fourier-based representations, whereas Chebyshev polynomial encoders address complex, non-linear, yet non-periodic relationships. This combination enhances the network's ability to learn diverse and explainable feature interactions without extensive manual feature engineering. Crucially, the resulting representations lend themselves to explanation: periodic encoders reveal the frequency components driving predictions, whereas Chebyshev encoders expose smooth, explainable transformations over non-repetitive domains.

This design directly addresses one of the main obstacles to industrial adoption of deep tabular models—their lack of inherent explainability. By embedding explainable inductive biases within the architecture, I shift explainability from a retrospective exercise to an intrinsic property of the model. The resulting framework allows practitioners not only to

evaluate feature contributions but also to interpret them in terms of domain-relevant cycles, trends, and deviations. This approach exemplifies the thesis's broader argument: to be epistemically robust, explainability must emerge from the structural consonance between the model and the world.

### 7.3.1 Contribution

This work introduces a novel family of deep architectures for explainable tabular learning grounded in periodic decomposition. I integrate Fourier-based encodings to capture cyclical regularities and Chebyshev polynomials to represent non-periodic dependencies, combining them in hybrid networks that adaptively balance both forms of structure. These architectures are designed to generalize across diverse datasets and task types—classification and regression alike—without prior feature categorization.

Empirical evaluations on a benchmark of 53 datasets demonstrate that this work's models outperform the FT-Transformer, the current SotA, on 34 datasets, confirming both their robustness and generality. Beyond raw performance, the proposed framework substantially enhances explainability: Fourier encodings expose the spectral composition of predictive reasoning, while Chebyshev transformations provide transparent mappings of complex feature interactions. By linking mathematical form to semantic meaning, this work's approach bridges DL and explainable modelling—offering an example of *structurally aware* AI in which explainability is not an afterthought but an architectural principle.

All code and datasets used in the experiments are publicly available to support reproducibility: `https://github.com/matteo-rizzo/periodic-tabular-dl`.

## 7.4 Related Work

Applying DL to tabular data has been a subject of significant research interest, aiming to replicate the success of DL in unstructured data domains such as computer vision and natural language processing [57]. This section reviews the literature on DL architectures tailored to tabular data, the use of Fourier transforms and Chebyshev polynomials in neural networks, and methods for capturing periodicity and nonlinearity in data.

### 7.4.1 Deep Learning Architectures for Tabular Data

Traditional ML models, particularly ensemble methods like Gradient Boosting Machines [53] and Random Forests [15], have historically outperformed DL models on tabular datasets [169]. These models are adept at handling heterogeneous feature types and complex interactions without extensive preprocessing. However, DL offers potential advantages in automatic feature extraction and representation learning. TabNet introduced a sequential attention mechanism that enables the model to focus on the most relevant features at each decision step [8]. By mimicking the decision-making process of gradient boosting, TabNet integrates feature selection with explainability, thereby improving the handling of the unique characteristics of tabular data. Neural Oblivious Decision Ensembles (NODE) propose an architecture that integrates differentiable decision trees into a neural network framework [142]. NODE uses oblivious decision trees, in which the same feature and threshold are applied at all decision nodes at a given depth, thereby enabling efficient representation learning and scalability. DeepGBM combined the strengths of gradient boosting and deep neural networks by using gradient boosting trees to preprocess the data

and generate input features for the neural network [91]. This hybrid approach leverages the powerful feature transformations of gradient boosting while benefiting from the representation learning capabilities of DL. FT-Transformer applied the Transformer architecture to tabular data by utilizing feature tokenization and self-attention mechanisms [58]. By treating each feature as a token, FT-Transformer models capture feature interactions through self-attention, enabling the detection of both linear and non-linear relationships without requiring explicit feature engineering. TabTransformer focused on modelling categorical features in tabular data using Transformer-based embeddings and self-attention [80]. TabTransformer captures dependencies and interactions that traditional one-hot encoding methods might miss by learning contextual embeddings for categorical variables. Despite these advancements, challenges remain in modelling the complex and diverse patterns inherent in tabular data, particularly in capturing periodicity and nonlinearity without substantial manual feature engineering.

### 7.4.2 Capturing Periodicity with Fourier Transforms

Periodicity is a common characteristic in various data types, manifesting as repeating patterns that recur at regular intervals. Traditional methods for handling periodicity often rely on feature engineering, such as creating lag features or applying domain-specific transformations [11]. Fourier transforms have been employed in neural networks to capture periodic patterns by transforming data into the frequency domain [202]. Fourier features allow models to represent functions with high-frequency components more effectively. Random Fourier Features were introduced to efficiently approximate kernel functions [145]. By mapping input data into a randomized low-dimensional feature space using sinusoidal functions, these features enable linear models to capture non-linear patterns associated with periodicity. Positional encoding in Transformers utilizes sinusoidal functions to inject sequence information into sequential data models [181]. The sine and cosine functions of varying frequencies enable the model to distinguish between different positions in the input sequence, implicitly capturing periodic relationships. Implicit neural representations with periodic activation functions have been explored to model high-frequency variations in data [170]. Using sinusoidal activation functions, neural networks can represent detailed signals and textures, which is particularly useful for tasks such as image generation and reconstruction. However, the application of Fourier-based encoding to general tabular data, which may not have explicit temporal or spatial dimensions, has been limited. This study extends Fourier transforms to tabular data by designing a Fourier-based neural encoder that captures intrinsic periodic patterns within the features without relying on explicit time or spatial information.

### 7.4.3 Modeling Nonlinearity with Chebyshev Polynomials

Non-linear relationships are prevalent in tabular data and challenge models that primarily capture linear interactions. Chebyshev polynomials, a sequence of orthogonal polynomials, are well-suited for approximating complex non-linear functions due to their minimax properties and numerical stability [148]. In neural networks, Chebyshev polynomials have been utilized in several contexts. A recent study [172] presents the Chebyshev Kolmogorov-Arnold Network, a novel neural network architecture inspired by the Kolmogorov-Arnold representation theorem that leverages the robust approximation capabilities of Chebyshev polynomials. This is achieved by employing learnable functions,

which are parameterized by Chebyshev polynomials, along the network's edge. Spectral Graph Convolutional Networks have employed Chebyshev polynomials to define convolutional filters in the spectral domain of graphs [36]. By approximating the graph Laplacian's eigenvalues, these models perform localized filtering operations without needing explicit eigenvalue decomposition. For function approximation, Chebyshev polynomials have been used to approximate arbitrary functions within neural networks [177]. This approach allows networks to represent functions with rapid variations and nonlinearities efficiently. The proposed Chebyshev-based neural encoder leverages these properties to capture non-periodic, complex, non-linear patterns in tabular data. By transforming input features through Chebyshev polynomials, the encoder enables the neural network to approximate intricate relationships that standard linear or non-linear transformations might not capture effectively.

### 7.4.4 Integrated Approaches for Periodic and Non-Periodic Patterns

While Fourier transforms and Chebyshev polynomials individually address periodicity and nonlinearity, real-world tabular datasets often exhibit both. Integrated approaches are necessary to capture the full spectrum of relationships within the data. The proposed architectures, PNPNet and AutoPNPNet, combine the strengths of both Fourier and Chebyshev encoders. PNPNet involves an *a priori* separation of features into periodic and non-periodic categories. This explicit division allows each encoder to specialize in modelling its designated feature type, improving the overall representation learning. AutoPNPNet addresses the limitations of manual feature separation by feeding all features into both encoders. An attention mechanism learns to automatically weigh and select features from each encoder, effectively performing feature selection and combination in a data-driven manner. These architectures draw inspiration from models that incorporate multiple feature transformations or branches to capture different aspects of the data [26] [181]. By integrating specialized encoders, the models can handle heterogeneous patterns more effectively than architectures that rely on a single transformation method.

### 7.4.5 Feature Selection and Automatic Relevance Determination

Feature selection is crucial in modelling tabular data because irrelevant or redundant features can degrade model performance. Neural networks often lack inherent mechanisms for feature selection, prompting research into integrating feature selection into DL models [27]. Attention mechanisms have been widely used to enable models to focus on the most relevant parts of the input [10, 139]. In tabular data models, attention can be applied to features, enabling the network to assign dynamic weights to each feature. Sparse regularization techniques, such as L1 regularization, encourage sparsity in the model weights, effectively zeroing out less important features [175]. This approach can lead to more explainable models and reduce overfitting. In the study by Li et al. [106], automatic feature selection networks have been proposed to learn feature importance scores during training. These networks can suppress irrelevant features and enhance the representation of significant ones, improving both performance and explainability. TabNet [8] uses a sparse feature mask at each decision step to select important features on an instance-by-instance basis. This mask is trained with information from the previous step. A feature transformer module decides which features to use for current prediction and which to pass to the next step. Some transformer layers are shared across steps, and the combined feature masks

provide global feature importance scores. A recent study by Amballa et al. [7] automated and accelerated feature selection using Priority-Based Random Grid Search and Greedy Search. The Priority-Based method utilizes prior probabilities to sample and evaluate a limited number of feature combinations, thereby avoiding exhaustive testing. Greedy Search methods (Backward Elimination and Forward Selection) iteratively refine the feature set by adding or removing features based on their statistical significance. These methods reduce computational cost while maintaining high model performance by modeling feature interactions and evaluating subsets using metrics.

The later-detailed AutoPNPNet incorporates an attention mechanism that automatically selects and combines features from both the Fourier and Chebyshev encoders. This approach aligns with automatic relevance determination, where the model learns the importance of each feature or transformation without manual intervention [114].

### 7.4.6 Explainability for Tabular Deep Learning

With the rapid advancement of ML models, there has been a growing need to elucidate the autonomous decisions and actions of these models to human users. This need for transparency is critical to foster trust and understanding among users who rely on these systems for various applications. The widespread adoption of DL techniques, particularly for tabular data, has exacerbated the challenge of model explainability. These sophisticated models often function as "black boxes," providing limited insight into their decision-making processes. Consequently, the opacity of these models poses a substantial barrier to their acceptance and effective utilization, as users are left with inadequate explanations for the outcomes generated by these systems.

In tabular DL, explainability faces unique challenges arising from the structured nature of the data and the representations learned by deep models. However, the opacity of these models has necessitated the application of dedicated XAI techniques for tabular data [133]. A recent study by O'Brien et al. [133] reviews XAI techniques for tabular data, drawing upon prior work, particularly a survey of explainable artificial intelligence for tabular data, and explores recent developments. This study classifies and outlines XAI methods pertinent to tabular data, highlights domain-specific challenges and gaps, and investigates potential applications and emerging trends. They provide an up-to-date overview of XAI techniques for tabular data, categorizing and describing methods, identifying domain-specific challenges, and exploring potential applications and trends in this field. Recent advances have introduced deep architectures that are inherently explainable and tailored to tabular data. For instance, InterpreTabNet improves classification accuracy and explainability by leveraging the TabNet architecture with an improved attentive module, thereby ensuring robust gradient propagation and computational stability [186].

Another significant contribution by Ullah et al. [178] introduces Layer-wise Relevance Propagation, an explainability method applied to tabular datasets using DL models. This method has been utilized for credit card fraud detection and predicting telecom customer churn. The growing demand for transparency in healthcare and other critical sectors has further fueled scholarly interest in these models. For instance, patient diagnosis can be achieved using tabular data from patient records. The study by [190] provides an in-depth analysis of recent research and advances in XAI, as well as its applications in the Internet of Medical Things within healthcare facilities.

Explainable tabular data analysis is also pertinent in the financial sector [191]. Amballa et al. [19] highlight the significant role of XAI in the financial industry, particularly

in applications of risk management, which include fraud detection, loan default prediction, and bankruptcy prediction. In industrial manufacturing, XAI-based tabular learning is particularly valuable for quality control [99].

### 7.4.7 Summary and Positioning

The existing literature highlights the challenges and potential solutions for applying DL to tabular data. While previous models have introduced innovative architectures and mechanisms to handle feature heterogeneity and interactions, a gap remains in effectively capturing intrinsic periodicity and complex nonlinear patterns without relying on extensive feature engineering. This work differentiates itself by leveraging intrinsic periodicity through a Fourier-based neural encoder specifically designed to capture periodic patterns in tabular data, extending the application of Fourier transforms beyond domains with explicit temporal or spatial structures. I model complex nonlinearities using Chebyshev polynomials in a neural encoder to approximate nonlinear functions on tabular data, providing a powerful tool for representing intricate relationships. By integrating specialized encoders, I develop architectures that combine both encoders to capture a wide range of patterns with automatic feature selection and combination mechanisms. Through empirical validation, I demonstrate the effectiveness of the approaches in this work through extensive experiments on diverse datasets, yielding significant performance improvements over existing DL methods. This work's method advances DL for tabular data by addressing limitations of current models and introducing novel methods to capture periodicity and nonlinearity. It opens avenues for further research into specialized encoders and integrated architectures that address the unique challenges posed by structured datasets.

### 7.4.8 Challenges with Learning in Tabular Data

- **Tabular data availability and generation:** Collecting, encoding, synthesizing, generating, and evaluating tabular data is challenging because of its diverse nature, complex patterns, and the absence of standard benchmarks [188]. Additionally, tabular data include categorical variables that should be transformed into numerical types for use by DL algorithms [68].

- **Preprocessing of tabular data:** DL applications with homogeneous data, only minimal preprocessing or explicit feature engineering is required, but using deep neural networks, especially while using tabular data, needs a specific strategy to apply preprocessing techniques [59]. Preprocessing techniques for deep neural networks can introduce information loss, which may reduce predictive performance [50].

- **Feature engineering problems:** Traditional DL architectures are not well-suited for handling the heterogeneous features of tabular data. There are challenges in effectively embedding both categorical and numerical features, often requiring manual or hybrid techniques to optimize performance. The lack of inherent spatial structure in tabular data exacerbates these issues [196].

- **Lack of inherent explainability:** DL models, particularly tabular data models, are likely to be "black boxes" and therefore are difficult to interpret their decision process. Deep neural networks are opaque, unlike other ML models, such as decision trees or linear regression, whose architectures are transparent and explainable. This

lack of explainability is a concern in critical applications such as healthcare and finance, where model predictions must be interpretable to support trust and compliance.

## 7.5 Method

This section discusses the proposed neural architectures for processing tabular data in this work. These are FourierNet, ChebyshevNet, PNPNet, AutoPNPNet, which process only continuous numerical data, and TabFourierNet, TabChebyshevNet, TabPNPNet, and TabAutoPNPNet, which process both continuous numerical and categorical features. FourierNet is a neural network architecture that incorporates a Fourier-based encoder. The Fourier encoder transforms input features into a frequency domain representation, enabling the network to capture periodic patterns effectively. By applying the Fourier Transform, I decompose complex periodic signals into their constituent sinusoidal components, thereby enabling the model to learn from frequency-based features that may be obscured in the original domain. However, not all patterns in tabular data are periodic. Datasets may also contain non-periodic, complex, and non-linear relationships that are crucial for accurate predictions. To capture these patterns, I developed ChebyshevNet, which employs a Chebyshev-based neural encoder. Chebyshev polynomials are a sequence of orthogonal polynomials that approximate functions on a specific interval and are particularly effective for modelling non-linear behaviour. Using Chebyshev polynomials, I can approximate complex functions and capture intricate relationships within the data without assuming periodicity. FourierNet and ChebyshevNet process only continuous data. TabFourierNet and TabChebyshevNet also incorporate categorical data by processing it on a separate branch. Categorical data is one-hot encoded and passed through a Multi Layer Perceptron (MLP) for feature extraction. Then, the extracted features are concatenated with those extracted in the parallel branch for continuous data. Figure 7.1 provides a general working schema for TabFourierNet and TabChebyshevNet, illustrating their shared underlying mechanism.

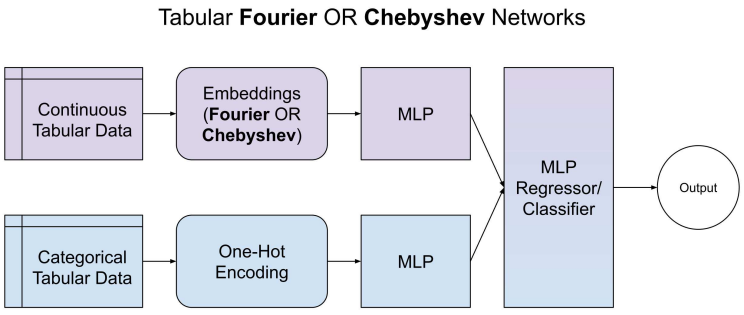



Figure 7.1: Shared base architecture of TabFourierNet and TabChebyshevNet.

Building upon these specialized encoders, I propose two integrated architectures: PNPNet and AutoPNPNet. The PNPNet architecture involves an a priori separation of input

features into periodic and non-periodic categories based on domain knowledge or statistical analysis. The periodic features are processed through the Fourier encoder, while the non-periodic features pass through the Chebyshev encoder. The outputs from both branches are then combined to produce the final prediction. This explicit separation allows the model to tailor its representation learning to the nature of each feature type. Recognizing that manual feature separation may not always be feasible or accurate, AutoPNPNet feeds all features into the Fourier and Chebyshev encoders. An additional MLP layer learns to automatically weigh and select the most relevant features from each encoder, performing implicit feature selection and combination. This approach eliminates the need for prior feature categorization, allowing the model to learn the optimal representation adaptively. From PNPNet and AutoPNPNet, I derive TabPNPNet and TabAutoPNPNet that process categorical data on a separate branch from continuous data. Figure 7.2 and Figure 7.3 present the architecture of PNPNet and AutoPNPNet, respectively, highlighting how periodic and non-periodic feature encodings are integrated.

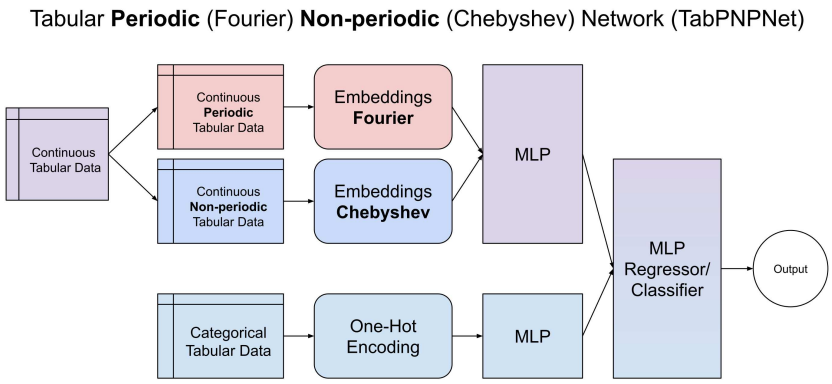



Figure 7.2: Overview of the PNPNet architecture.

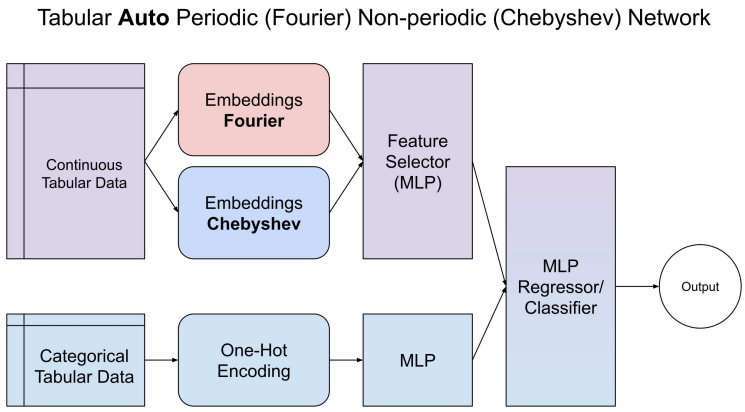



Figure 7.3: Overview of the AutoPNPNet architecture.

### 7.5.1 FourierNet: Capturing Periodic Patterns

The Fourier-based neural encoder transforms input features into a frequency-domain representation, enabling the network to learn effectively from periodic components in the data. This highly configurable encoder offers a range of options, including input scaling, convolutional preprocessing, different activation functions, learnable frequencies, phase shifts, amplitudes, and Random Fourier Features (RFF). Given an input feature vector $\mathbf{x} \in \mathbb{R}^d$, the encoder processes $\mathbf{x}$ through several optional steps before applying the Fourier transformation. If input scaling is enabled, each input feature is scaled by a learnable parameter $\mathbf{s} \in \mathbb{R}^d$:

$$\tilde{\mathbf{x}} = \mathbf{s} \odot \mathbf{x},$$

where $\odot$ denotes element-wise multiplication. This scaling allows the model to adjust the influence of each feature. An optional one-dimensional convolutional layer can be applied to capture localized feature patterns:

$$\tilde{\mathbf{x}} = \text{Conv1D}(\tilde{\mathbf{x}}),$$

where Conv1D represents a convolutional operation over the feature dimension with a specified kernel size.

The core of the Fourier encoder is to project the input features into a higher-dimensional space using sinusoidal functions. For each input feature $x_i$, I associate $m$ frequency components, forming a frequency matrix $\mathbf{F} \in \mathbb{R}^{d \times m}$. Frequencies $f_{ij}$ can be initialized using uniform, normal, or logarithmic initialization methods. If RFF is enabled, frequencies are sampled from a normal distribution with variance inversely proportional to the bandwidth parameter $\sigma$:

$$f_{ij} \sim \mathcal{N}\left(0, \frac{1}{\sigma^2}\right).$$

Optionally, the encoder includes learnable amplitude $\alpha_{ij}$ and phase shift $\phi_{ij}$ parameters for each feature and frequency component. The amplitude scaling is given by $a_{ij} = \alpha_{ij}$, and the phase shift is $\varphi_{ij} = \phi_{ij}$. The input features are projected using the frequencies, amplitudes, and phase shifts as:

$$z_{ij} = 2\pi x_i f_{ij} + \varphi_{ij}.$$

The encoder supports several activation functions applied to projected inputs $z_{ij}$. For instance, with the sine and cosine activation ('`sin_cos`'), the outputs are:

$$F_{ij}^{\sin} = a_{ij} \sin(z_{ij}), \quad F_{ij}^{\cos} = a_{ij} \cos(z_{ij}).$$

These outputs are concatenated along the feature dimension. After activation, an optional learnable scaling parameter $\gamma_{ij}$ can be applied to encoded features:

$$F_{ij} = \gamma_{ij} F_{ij}.$$

The encoded features are then flattened to form the final encoded vector $\mathbf{F}(\mathbf{x})$ of dimension $d \times m \times s$, where $s$ depends on the activation function (e.g., $s = 2$ for '`sin_cos`', $s = 1$ otherwise):

$$\mathbf{F}(\mathbf{x}) = \text{Flatten}\left(F_{ij}\right).$$

The overall architecture of FourierNet integrates the Fourier encoder with an MLP. The Fourier encoder transforms the input features into a higher-dimensional encoded representation $\mathbf{F}(\mathbf{x})$ that captures periodic patterns via sinusoidal transformations with configurable frequencies, amplitudes, and phase shifts. This encoded representation is then processed by the MLP, which consists of $L$ hidden layers. The operations of the MLP are defined as:

$$\begin{aligned}\mathbf{h}_0 &= \mathbf{F}(\mathbf{x}),\\ \mathbf{h}_l &= \sigma\left(\mathbf{W}_l\mathbf{h}_{l-1} + \mathbf{b}_l\right), \quad \text{for } l = 1, 2, \ldots, L,\\ \hat{y} &= f_{\text{out}}(\mathbf{h}_L),\end{aligned}$$

where $\sigma(\cdot)$ is an activation function (e.g., ReLU), and $f_{\text{out}}(\cdot)$ is appropriate for the task (e.g., linear activation for regression, softmax for classification).

Combining all components, the encoded feature for each input feature $x_i$ and frequency component $j$ is given by:

$$F_{ij} = \gamma_{ij} \cdot \alpha_{ij} \cdot \sin\left(2\pi x_i f_{ij} + \phi_{ij}\right),$$

or according to the selected activation function. If the '`sin_cos`' activation is used, the encoded features are:

$$\begin{aligned}F_{ij}^{\text{sin}} &= \gamma_{ij} \cdot \alpha_{ij} \cdot \sin\left(2\pi x_i f_{ij} + \phi_{ij}\right),\\ F_{ij}^{\text{cos}} &= \gamma_{ij} \cdot \alpha_{ij} \cdot \cos\left(2\pi x_i f_{ij} + \phi_{ij}\right).\end{aligned}$$

These features are then concatenated and flattened to form $\mathbf{F}(\mathbf{x})$.

### 7.5.2 ChebyshevNet: Modeling Non-Periodic Patterns

To capture non-periodic, complex, non-linear relationships in tabular data, ChebyshevNet utilizes Chebyshev polynomials of the first kind. For each input feature $x_i$, the Chebyshev polynomials $T_n(x_i)$ are defined recursively:

$$T_0(x_i) = 1, \tag{7.1}$$
$$T_1(x_i) = x_i, \tag{7.2}$$
$$T_n(x_i) = 2x_i T_{n-1}(x_i) - T_{n-2}(x_i), \quad \text{for } n \geq 2. \tag{7.3}$$

A preliminary analysis of training trends revealed significant fluctuations in loss values, which hindered learning, particularly for higher-degree polynomials, which are more susceptible to numerical overflow. To address this issue, I explored feature clamping as a method to enhance training stability. The input features are clamped to the interval $[-1, 1]$:

$$x_i \leftarrow \text{clip}(x_i, -1, 1). \tag{7.4}$$

The study on how this choice affects performance showed no significant change in accuracy for lower-degree polynomials, whereas it yielded superior training stability for higher-degree polynomials.

The Chebyshev encoder $\mathbf{C}(\mathbf{x})$ generates an encoded feature tensor by computing the Chebyshev polynomials up to degree $N$ for each input feature, resulting in a tensor of shape $[d, N+1]$, where $d$ is the number of input features:

$$\mathbf{C}(\mathbf{x}) = [T_0(x_i),\ T_1(x_i),\ \ldots,\ T_N(x_i)]_{i=1}^{d}\,. \tag{7.5}$$

A multi-headed encoding mechanism is introduced to enhance the encoder's representative capacity. Let there be $H$ heads in total. Each head processes the input features independently, with its own set of learnable parameters, including optional scaling factors $\mathbf{s}^{(h)} \in \mathbb{R}^d$, polynomial weights $\mathbf{W}^{(h)} \in \mathbb{R}^{d\times(N+1)}$, and interaction kernels $\mathbf{K}^{(h)} \in \mathbb{R}^{d\times(N+1)\times q}$, where $q$ is the kernel size.

For each head $h$, the input features may be scaled:

$$\tilde{x}_i^{(h)} = s_i^{(h)} x_i. \tag{7.6}$$

The Chebyshev polynomials are then computed for each scaled feature $\tilde{x}_i^{(h)}$ up to degree $N$. The weighted polynomials are obtained by applying the polynomial weights:

$$P_{i,n}^{(h)} = W_{i,n}^{(h)} T_n(\tilde{x}_i^{(h)}). \tag{7.7}$$

To model interactions among polynomial terms, the kernels are applied:

$$S_{i,j}^{(h)} = \sum_{n=0}^{N} P_{i,n}^{(h)} K_{i,n,j}^{(h)}, \quad \text{for } j = 1, 2, \ldots, k. \tag{7.8}$$

An activation function $\sigma(\cdot)$ (e.g., SiLU) is applied to the result:

$$S_{i,j}^{(h)} = \sigma(S_{i,j}^{(h)}). \tag{7.9}$$

If residual connections are enabled, a residual from the input feature is added:

$$S_{i,j}^{(h)} = S_{i,j}^{(h)} + \tilde{x}_i^{(h)}. \tag{7.10}$$

Each head produces an output tensor $\mathbf{S}^{(h)} \in \mathbb{R}^{d\times k}$, representing the encoded features from that head. A cross-head attention mechanism is incorporated to capture dependencies between different heads. The output tensors from each head are flattened to form vectors:

$$\mathbf{s}^{(h)} = \text{Flatten}\left(\mathbf{S}^{(h)}\right), \quad \mathbf{s}^{(h)} \in \mathbb{R}^{dk}. \tag{7.11}$$

Queries, keys, and values for each head are computed using linear projections:

$$\mathbf{q}^{(h)} = \mathbf{W}_q \mathbf{s}^{(h)}, \tag{7.12}$$

$$\mathbf{k}^{(h)} = \mathbf{W}_k \mathbf{s}^{(h)}, \tag{7.13}$$

$$\mathbf{v}^{(h)} = \mathbf{W}_v \mathbf{s}^{(h)}, \tag{7.14}$$

where $\mathbf{W}_q, \mathbf{W}_k, \mathbf{W}_v \in \mathbb{R}^{d_s\times d_a}$ are learnable projection matrices, and $d_a$ is the attention dimension. The attention scores between heads are computed as:

$$\alpha^{(h,h')} = \frac{\left(\mathbf{q}^{(h)}\right)^\top \mathbf{k}^{(h')}}{\sqrt{d_a}}. \tag{7.15}$$

The attention weights are obtained via the softmax function:

$$a^{(h,h')} = \frac{\exp\left(\alpha^{(h,h')}\right)}{\sum_{h''=1}^{H} \exp\left(\alpha^{(h,h'')}\right)}. \tag{7.16}$$

The output of the cross-head attention for each head is calculated as:

$$\mathbf{o}^{(h)} = \sum_{h'=1}^{H} a^{(h,h')} \mathbf{v}^{(h')}. \tag{7.17}$$

The outputs from all heads are concatenated to form the final encoded representation:

$$\mathbf{C}(\mathbf{x}) = \text{Concat}\left(\mathbf{o}^{(1)}, \mathbf{o}^{(2)}, \ldots, \mathbf{o}^{(H)}\right). \tag{7.18}$$

An optional normalization layer, such as Layer Normalization, may be applied to $\mathbf{C}(\mathbf{x})$ to stabilize training.

ChebyshevNet integrates the Chebyshev encoder with the multi-headed encoding and cross-head attention mechanisms. The overall architecture begins with input processing, where input features may be scaled and clamped to $[-1, 1]$ for numerical stability. The multi-headed encoding then computes the weighted Chebyshev polynomials and applies the kernels for each head, including activation functions and residual connections as configured. Following encoding, the cross-head attention mechanism processes the outputs of all heads, computing queries, keys, values, and attention weights to combine the value vectors. The attended outputs of all heads are concatenated, and optional normalization is applied to the combined representation. The final encoded representation $\mathbf{C}(\mathbf{x})$ is passed through an MLP:

$$\mathbf{h}_0 = \mathbf{C}(\mathbf{x}), \tag{7.19}$$

$$\mathbf{h}_l = \sigma\left(\mathbf{W}_l \mathbf{h}_{l-1} + \mathbf{b}_l\right), \quad \text{for } l = 1, 2, \ldots, L, \tag{7.20}$$

$$\hat{y} = f_{\text{out}}\left(\mathbf{h}_L\right). \tag{7.21}$$

Here, $\sigma(\cdot)$ is an activation function (e.g., ReLU), and $f_{\text{out}}(\cdot)$ is the output function appropriate for the task.

### 7.5.3 PNPNet: Periodic-Non-Periodic Network

PNPNet requires an initial separation of the input features into periodic and non-periodic subsets, denoted as $\mathbf{x}_p \in \mathbb{R}^{d_p}$ and $\mathbf{x}_{np} \in \mathbb{R}^{d_{np}}$, respectively, where $d_p + d_{np} = d$. This separation can be based on domain knowledge or statistical analysis techniques, such as spectral density estimation or autocorrelation analysis.

PNPNet comprises two parallel branches that process periodic and non-periodic features separately. The periodic features $\mathbf{x}_p$ are transformed using the Fourier encoder to obtain $\mathbf{F}(\mathbf{x}_p)$, which is then processed by an MLP specific to the periodic branch:

$$\mathbf{h}_{p,0} = \mathbf{F}(\mathbf{x}_p), \tag{7.22}$$

$$\mathbf{h}_{p,l} = \sigma\left(\mathbf{W}_{p,l} \mathbf{h}_{p,l-1} + \mathbf{b}_{p,l}\right), \quad \text{for } l = 1, 2, \ldots, L_p. \tag{7.23}$$

Similarly, the non-periodic features $\mathbf{x}_{np}$ are transformed using the Chebyshev encoder to obtain $\mathbf{C}(\mathbf{x}_{np})$, processed by the non-periodic branch MLP:

$$\mathbf{h}_{np,0} = \mathbf{C}(\mathbf{x}_{np}), \tag{7.24}$$

$$\mathbf{h}_{np,l} = \sigma\left(\mathbf{W}_{np,l} \mathbf{h}_{np,l-1} + \mathbf{b}_{np,l}\right), \quad \text{for } l = 1, 2, \ldots, L_{np}. \tag{7.25}$$

The outputs from both branches are concatenated to form the fused representation:

$$\mathbf{h}_{\text{fusion}} = \begin{bmatrix} \mathbf{h}_{p,L_p} \\ \mathbf{h}_{np,L_{np}} \end{bmatrix}. \tag{7.26}$$

This fused representation is then passed through a final MLP to produce the prediction:

$$\mathbf{h}_{\text{final}} = \sigma\left(\mathbf{W}_f \mathbf{h}_{\text{fusion}} + \mathbf{b}_f\right), \tag{7.27}$$

$$\hat{y} = f_{\text{out}}\left(\mathbf{h}_{\text{final}}\right). \tag{7.28}$$

### 7.5.4 AutoPNPNet: Automatic Feature Selection

AutoPNPNet addresses the challenge of manually separating features into periodic and non-periodic categories by simultaneously feeding all input features $\mathbf{x} \in \mathbb{R}^d$ into the Fourier and Chebyshev encoders. This approach enables the model to learn, in a data-driven manner, which features are best represented by each encoder. The input features are processed as:

$$\mathbf{F}(\mathbf{x}) = \text{FourierEncoder}(\mathbf{x}), \tag{7.29}$$

$$\mathbf{C}(\mathbf{x}) = \text{ChebyshevEncoder}(\mathbf{x}). \tag{7.30}$$

Here, $\mathbf{F}(\mathbf{x}) \in \mathbb{R}^{d_F}$ and $\mathbf{C}(\mathbf{x}) \in \mathbb{R}^{d_C}$ are the encoded representations produced by the Fourier and Chebyshev encoders, respectively, where $d_F$ and $d_C$ are the dimensions of the encoded feature spaces.

The encoded representations from both encoders are processed through their respective MLPs to capture complex interactions and feature transformations:

$$\mathbf{h}_{F,0} = \mathbf{F}(\mathbf{x}), \tag{7.31}$$

$$\mathbf{h}_{F,l} = \sigma\left(\mathbf{W}_{F,l}\mathbf{h}_{F,l-1} + \mathbf{b}_{F,l}\right), \quad l = 1, 2, \ldots, L_F, \tag{7.32}$$

$$\mathbf{h}_{C,0} = \mathbf{C}(\mathbf{x}), \tag{7.33}$$

$$\mathbf{h}_{C,l} = \sigma\left(\mathbf{W}_{C,l}\mathbf{h}_{C,l-1} + \mathbf{b}_{C,l}\right), \quad l = 1, 2, \ldots, L_C. \tag{7.34}$$

Here, $\mathbf{h}_{F,L_F} \in \mathbb{R}^{h_F}$ and $\mathbf{h}_{C,L_C} \in \mathbb{R}^{h_C}$ are the outputs of the Fourier and Chebyshev branches after the $L_F$ and $L_C$ layers, respectively.

An attention mechanism is employed to learn the optimal weighting between the two branches, allowing the model to automatically determine the importance of periodic and non-periodic features. The attention weights $\alpha_F$ and $\alpha_C$ are computed as:

$$\boldsymbol{\alpha} = \text{softmax}\left(\mathbf{W}_{\text{att}} \begin{bmatrix} \mathbf{h}_{F,L_F} \\ \mathbf{h}_{C,L_C} \end{bmatrix} + \mathbf{b}_{\text{att}}\right), \tag{7.35}$$

$$\boldsymbol{\alpha} = \begin{bmatrix} \alpha_F \\ \alpha_C \end{bmatrix}. \tag{7.36}$$

where $\mathbf{W}_{\text{att}} \in \mathbb{R}^{2\times(h_F+h_C)}$ and $\mathbf{b}_{\text{att}} \in \mathbb{R}^2$ are learnable parameters, and the softmax function ensures that $\alpha_F + \alpha_C = 1$ and $\alpha_F, \alpha_C \geq 0$.

The fused representation is then computed as a weighted sum of the outputs from the two branches:

$$\mathbf{h}_{\text{fusion}} = \alpha_F \mathbf{h}_{F,L_F} + \alpha_C \mathbf{h}_{C,L_C}. \tag{7.37}$$

This fusion enables the model to emphasize the features and patterns most relevant to the prediction task, as determined by the learned attention weights.

The fused representation $\mathbf{h}_{\text{fusion}}$ is then passed through a final MLP to produce the prediction $\hat{y}$:

$$\mathbf{h}_{\text{final}} = \sigma\left(\mathbf{W}_f \mathbf{h}_{\text{fusion}} + \mathbf{b}_f\right), \ \hat{y} \qquad = f_{\text{out}}\left(\mathbf{h}_{\text{final}}\right). \tag{7.38}$$

Here, $\mathbf{W}_f \in \mathbb{R}^{h\text{final} \times h_{\text{fusion}}}$ and $\mathbf{b}_f \in \mathbb{R}^{h\text{final}}$ are the weights and biases of the final MLP layer, $\sigma(\cdot)$ is an activation function (e.g., ReLU), and $f_{\text{out}}(\cdot)$ is the output function appropriate for the task (e.g., linear activation for regression, softmax for classification).

In certain configurations, AutoPNPNet may incorporate a cross-attention mechanism to capture additional interactions between the features encoded by the Fourier and Chebyshev encoders. The cross-attention enables the model to attend to the combined representations, thereby enhancing its ability to model complex dependencies.

Given the projected representations $\mathbf{z}_F$ and $\mathbf{z}_C$ obtained from the encoded features:

$$\mathbf{z}_F = \mathbf{W}_{F,\text{proj}} \mathbf{h}_{F,L_F}, \ \mathbf{z}_C \qquad = \mathbf{W}_{C,\text{proj}} \mathbf{h}_{C,L_C}, \tag{7.39}$$

where $\mathbf{W}_{F,\text{proj}} \in \mathbb{R}^{d_{\text{embed}} \times h_F}$ and $\mathbf{W}_{C,\text{proj}} \in \mathbb{R}^{d_{\text{embed}} \times h_C}$ are projection matrices that correspond to a shared embedding space of dimension $d_{\text{embed}}$.

The combined representations are concatenated and passed through a multi-head attention layer:

$$\mathbf{H} = \text{MultiHeadAttention}\left([\mathbf{z}_F; \ \mathbf{z}_C]\right), \tag{7.40}$$

where $\mathbf{H} \in \mathbb{R}^{2 \times d_{\text{embed}}}$ captures the attended features, and the multi-head attention mechanism allows the model to focus on different aspects of the combined representations.

The attended features are then processed through additional layers, including a linear transformation, activation function, and normalization, as follows:

$$\mathbf{H}' = \sigma\left(\mathbf{W}_{\text{linear}} \mathbf{H} + \mathbf{b}_{\text{linear}}\right), \ \mathbf{H}'' \qquad = \text{LayerNorm}\left(\mathbf{H}'\right). \tag{7.41}$$

An aggregation operation, such as averaging, is applied across the sequence dimension to obtain a fixed-size representation:

$$\mathbf{h}_{\text{attn}} = \text{MeanPooling}\left(\mathbf{H}''\right). \tag{7.42}$$

This representation $\mathbf{h}_{\text{attn}}$ is then used in place of or in addition to $\mathbf{h}_{\text{fusion}}$ for the final prediction layer.

### 7.5.5 Training Objective

The training objective depends on the task.

For regression, I minimize the Mean Squared Error (MSE) loss:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \tag{7.43}$$

where $n$ is the number of samples, $y_i$ is the true target, and $\hat{y}_i$ is the predicted value.

For classification, I minimize the Cross-Entropy loss:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\log \hat{y}_{ik}, \tag{7.44}$$

where $K$ is the number of classes, $y_{ik}$ is the binary indicator (1 if sample $i$ belongs to class $k$, 0 otherwise), and $\hat{y}_{ik}$ is the predicted probability that sample $i$ belongs to class $k$.

### 7.5.6 Implementation Details

I use the Rectified Linear Unit (ReLU) activation function defined as:

$$\sigma(z) = \max(0, z). \tag{7.45}$$

Input features are normalized to have zero mean and unit variance. I employ batch normalization after each fully connected layer to improve training stability and convergence. I employ the Adam optimizer [96] with default parameters for training. The update rule for the parameters $\theta$ is:

$$\theta_{t+1} = \theta_t - \eta_t \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}, \tag{7.46}$$

where $\eta_t$ is the learning rate, $\hat{m}_t$ and $\hat{v}_t$ are the bias-corrected first and second moment estimates, and $\epsilon$ is a small constant to prevent division by zero. To avoid overfitting, I apply dropout regularization and L2 weight decay. Dropout randomly zeros a fraction of the neurons during training, reducing neuron co-adaptation. L2 regularization adds a penalty term proportional to the squared weights to the loss function.

### 7.5.7 Computational Complexity Analysis

The encoders introduce additional computational overhead compared to standard MLPs. The computational complexity of the Fourier encoder is $\mathcal{O}(d \cdot m)$, where $d$ is the number of input features and $m$ is the number of frequency components. The computational complexity of the Chebyshev encoder is $\mathcal{O}(d \cdot N)$, where $N$ is the maximum degree of the Chebyshev polynomials.

### 7.5.8 Periodicity Detection

Detecting periodic patterns in data is crucial for effectively modeling and forecasting time series and other datasets with cyclical behaviors. Accurate identification of periodicity allows models to incorporate appropriate transformations and encodings, such as Fourier

transforms, to capture these patterns. This section presents a method for detecting periodicity in time-series data using the autocorrelation function (ACF) and peak-detection algorithms. The ACF measures the correlation of a signal with a delayed copy of itself as a function of the delay (lag). For a discrete time series $\{x_t\}_{t=1}^{T}$, the autocorrelation at lag $k$ is defined as:

$$\rho_k = \frac{\sum_{t=1}^{T-k}(x_t - \bar{x})(x_{t+k} - \bar{x})}{\sum_{t=1}^{T}(x_t - \bar{x})^2}, \tag{7.47}$$

where $\bar{x}$ is the mean of the series, and $k$ is the lag. The ACF $\rho_k$ provides insight into the repeating patterns within the data by highlighting the lags in which the series is correlated with itself. Periodic time series exhibit significant autocorrelations at lags corresponding to their period and multiples. By analyzing the peaks in the ACF, I can infer periodicity and estimate the period length. To detect periodicity using the ACF, I first calculate the ACF of the time series up to a specified maximum lag $L_{\text{max}}$. The $L_{\text{max}}$ should cover at least one expected period. Next, I identify significant peaks in the ACF using a peak detection algorithm. Peaks represent lags where the autocorrelation is locally maximal and exceeds certain thresholds in height and prominence. Specifically, peaks must have a height above a minimum value $\rho_{\text{min}}$ to be considered significant, and they must stand out relative to neighbouring values, measured by the prominence parameter $p_{\text{min}}$. Additionally, peaks must be separated by at least a minimum number of lags $d_{\text{min}}$ to avoid closely spaced false positives. After identifying the peaks, I analyze the distances between consecutive peaks, denoted as $\Delta_k = \tau_{k+1} - \tau_k$, where $\tau_k$ is the lag of the $k$-th peak. If the series is periodic, these distances should be approximately constant, corresponding to the series period. I assess periodicity by determining whether the standard deviation of the peak distances $\Delta_k$ is below a certain threshold $\sigma_{\text{max}}$. A low standard deviation indicates that the peaks occur at regular intervals, supporting the presence of periodicity. Let $\{\rho_k\}_{k=1}^{L_{\text{max}}}$ be the autocorrelation values excluding lag zero. The peak detection algorithm identifies the set of peak lags $\{\tau_k\}$ satisfying the following conditions:

$$\begin{cases} \rho_{\tau_k} \geq \rho_{\text{min}}, \\ \text{Prominence}(\rho_{\tau_k}) \geq p_{\text{min}}, \\ \tau_k - \tau_{k-1} \geq d_{\text{min}}. \end{cases} \tag{7.48}$$

The standard deviation of the peak distances is computed as follows:

$$\sigma_\Delta = \sqrt{\frac{1}{N-1}\sum_{k=1}^{N-1}\left(\Delta_k - \bar{\Delta}\right)^2}, \tag{7.49}$$

where $N$ is the number of detected peaks and $\bar{\Delta}$ is the mean of the peak distances. Periodicity is detected if $\sigma_\Delta < \sigma_{\text{max}}$.

The following pseudocode summarizes the steps of the periodicity detection algorithm, which utilizes the ACF and peak detection.

**Algorithm 2:** Periodicity Detection using ACF Peaks

**Input:** Time series data $\{x_t\}_{t=1}^{T}$; Parameters: $L_{\max}$, $\rho_{\min}$, $p_{\min}$, $d_{\min}$, $\sigma_{\max}$
**Output:** Boolean value indicating whether periodicity is detected

Compute the ACF up to lag $L_{\max}$: $\{\rho_k\}_{k=0}^{L_{\max}} \leftarrow \text{ACF}(\{x_t\}, L_{\max})$;
Exclude lag zero: $\{\rho_k\}_{k=1}^{L_{\max}} \leftarrow \{\rho_k\}_{k=1}^{L_{\max}}$;
Initialize peak lags: $\mathcal{P} \leftarrow \emptyset$;
Initialize previous peak lag: $k_{\text{prev}} \leftarrow -d_{\min}$;
**for** $k = 1$ *to* $L_{max}$ **do**
    **if** $\rho_k \geq \rho_{min}$ *Prominence*$(\rho_k) \geq p_{min}$ $(k - k_{prev}) \geq d_{min}$ **then**
        Add $k$ to peak lags: $\mathcal{P} \leftarrow \mathcal{P} \cup \{k\}$;
        Update previous peak lag: $k_{\text{prev}} \leftarrow k$;
    **end**
**end**
Let $N$ be the number of detected peaks: $N \leftarrow |\mathcal{P}|$;
**if** $N \geq 2$ **then**
    Compute distances between consecutive peaks: $\Delta_k = \tau_{k+1} - \tau_k$, for $k = 1, \ldots, N-1$;
    Compute mean of distances: $\bar{\Delta} = \dfrac{1}{N-1} \sum_{k=1}^{N-1} \Delta_k$;
    Compute std dev of distances: $\sigma_\Delta = \sqrt{\dfrac{1}{N-1} \sum_{k=1}^{N-1} \left(\Delta_k - \bar{\Delta}\right)^2}$;
    **if** $\sigma_\Delta < \sigma_{max}$ **then**
        **return** *True* `// Periodicity detected`
    **else**
        **return** *False* `// Periodicity not detected`
    **end**
**else**
    **return** *False* `// Not enough peaks to assess periodicity`
**end**

## 7.6 Datasets

I utilize a benchmark of 53 tabular datasets encompassing regression and classification tasks [64]. These datasets span various domains and differ in sample size, feature types, and complexity, as shown in Tables 7.1, 7.2, 7.3, and 7.4. I categorize the datasets for analysis based on task type and feature composition. The task types include regression, which predicts continuous target variables, and classification, which predicts categorical target variables. The feature compositions include datasets with exclusively numerical features and those with a mix of numerical and categorical features. This categorization enables us to evaluate the models' performance under various data conditions.

I follow the data preprocessing protocols outlined in the benchmark study. Missing values are imputed using the mean for numerical features and the mode for categorical features. Categorical variables are processed using embedding layers within the models, consistent with FT-Transformer's methodology. The numerical features are standardized to have a mean of zero and a variance of one.

Table 7.1: Statistics of benchmark datasets for **numerical classification**.

| **Dataset** | **# Samples** | **# Features** |
|---|---|---|
| electricity | 38.474 | 7 |
| covertype | 566.602 | 10 |
| pol | 10.082 | 26 |
| house_16H | 13.488 | 16 |
| kdd_ipums_la_97-small | 5.188 | 20 |
| MagicTelescope | 13.376 | 10 |
| bank-marketing | 10.578 | 7 |
| phoneme | 3.172 | 5 |
| MiniBooNE | 72.998 | 50 |
| Higgs | 940.160 | 24 |
| eye_movements | 7.608 | 20 |
| jannis | 57.580 | 54 |
| credit | 16.714 | 10 |
| california | 20.634 | 8 |
| wine | 2.554 | 11 |

Table 7.2: Statistics of benchmark datasets for **numerical regression**.

| **Dataset** | **# Samples** | **# Features** |
|---|---|---|
| cpu_act | 8.192 | 21 |
| pol | 15.000 | 26 |
| elevators | 16.599 | 16 |
| isolet | 7.797 | 613 |
| wine_quality | 6.497 | 11 |
| Ailerons | 13.750 | 33 |
| houses | 20.640 | 8 |
| house_16H | 22.784 | 16 |
| diamonds | 53.940 | 6 |
| Brazilian_houses | 10.692 | 8 |
| Bike_Sharing_Demand | 17.379 | 6 |
| nyc-taxi-green-dec-2016 | 581.835 | 9 |
| house_sales | 21.613 | 15 |
| sulfur | 10.081 | 6 |
| medical_charges | 163.065 | 5 |
| MiamiHousing2016 | 13.932 | 14 |
| superconduct | 21.263 | 79 |
| california | 20.640 | 8 |
| fifa | 18.063 | 5 |
| year | 515.345 | 90 |

Table 7.3: Statistics of benchmark datasets for **mixed-feature classification**.

| Dataset | # Samples | # Features |
|---|---|---|
| electricity | 38.474 | 8 |
| eye_movements | 7.608 | 23 |
| KDDCup09_upselling | 5.032 | 45 |
| covertype | 423.680 | 54 |
| rl | 4.970 | 12 |
| road-safety | 111.762 | 32 |
| compas-two-years | 16.644 | 17 |

Table 7.4: Statistics of benchmark datasets for **mixed-feature regression**.

| Dataset | # Samples | # Features |
|---|---|---|
| yprop_4_1 | 8.885 | 62 |
| analcatdata_supreme | 4.052 | 7 |
| visualizing_soil | 8.641 | 4 |
| black_friday | 166.821 | 9 |
| diamonds | 53.940 | 9 |
| Mercedes_Benz_Greener_Manufacturing | 4.209 | 359 |
| Brazilian_houses | 10.692 | 11 |
| Bike_Sharing_Demand | 17.379 | 11 |
| OnlineNewsPopularity | 39.644 | 59 |
| nyc-taxi-green-dec-2016 | 581.835 | 16 |
| house_sales | 21.613 | 17 |
| particulate-matter-ukair-2017 | 394.299 | 6 |
| SGEMM_GPU_kernel_performance | 241.600 | 9 |

## 7.7 Experiments

The experiments conducted in this work are designed to address four key research questions:

- **RQ1**: Do FourierNet and ChebyshevNet (and respective Tab versions) individually outperform FT-Transformer on the benchmark?
- **RQ2**: Does integrating periodic and non-periodic encoders in PNPNet and AutoPNPNet (and respective Tab versions) lead to further performance gains compared to FT-Transformer?
- **RQ3**: How does AutoPNPNet's automatic feature selection mechanism compare to the manual feature separation employed in PNPNet?
- **RQ4**: What is the trade-off between the computational overhead introduced by the specialized encoders and the performance improvements achieved?

As discussed in § 7.6, I evaluate the performance of the proposed models on a comprehensive benchmark of tabular datasets, encompassing both regression and classifica-

tion tasks, with continuous numerical and categorical features. I evaluate the FourierNet, ChebyshevNet, PNPNet, and AutoPNPNet models on datasets with only continuous numerical features. Conversely, I evaluate the TabFourierNet, TabChebyshevNet, TabPNPNet, and TabAutoPNPNet models on datasets comprising both continuous numerical and categorical features. I compare the proposed models against FT-Transformer, the SotA baseline for DL method operating on tabular data[1]. I deliberately omit other DL baselines since FT-Transformer beat them on the examined benchmark, and other tree-based models, as they are far superior to deep models, as extensively discussed in Grinsztajn et al. [64].

To ensure robust evaluation, I perform 5-fold cross-validation on each dataset. All models, including the FT-Transformer and the proposed architectures in this work, are configured with comparable hyperparameters to ensure a fair comparison. Each model uses an MLP with four hidden layers, each with 256 neurons. The ReLU activation function is employed for hidden layers, linear activation for regression tasks, and softmax for classification tasks. The models are optimized using the Adam optimizer with a learning rate of 0.001. The batch size is set to 1,024 samples, and training proceeds for up to 100 epochs with early stopping based on validation loss, using a patience of 10. Regularization techniques include a dropout rate of 0.1 and an L2 weight decay coefficient of 0.0001. For this work's specialized encoders, the Fourier encoder uses 20 frequency components, with frequencies selected based on the range of the input features. The Chebyshev encoder has a maximum polynomial degree of 10. The attention mechanism in AutoPNPNet is implemented using a learnable gating mechanism.

I employ standard evaluation metrics suitable for regression and classification tasks. For regression tasks, I use Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the Coefficient of Determination ($R^2$ Score). I use Accuracy, Precision, Recall, and F1-Score for classification tasks. The tables in the following sections report the improved metrics for each task and dataset, the model that reported the improvement, and the percentage of improvement over the baseline FT-Transformer. Such an improvement percentage is computed by analyzing the mean and standard deviation of each metric across the five folds. A Wilcoxon signed-rank test (per metric) was used to compare each model with the baseline across datasets.

## 7.8 Results

### 7.8.1 Regression Tasks

Tables 7.5 and 7.6 present the performance improvements of this work's model for the mixed-feature and numerical regression tasks, respectively. The proposed models outperform FT-Transformer in at least one metric on 9 of 13 datasets containing both categorical and continuous numerical features and on 12 of 20 datasets with only continuous numerical features.

[1]This work's reference implementation of FT-Transformer is presented in Gorishniy et al. [58], and offers a straightforward open-source implementation available at: https://github.com/yandex-research/rtdl-revisiting-models

Table 7.5: Performance improvements of proposed models over FT-Transformer for **mixed-feature regression** tasks. Relative (%) improvements are reported for scale-dependent metrics (RMSE, MAE); absolute $\Delta R^2$ is shown for $R^2$.

| **Dataset** | **Metric** | **Best Model** | **Improvement** |
|---|---|---|---|
| Bike_Sharing_Demand | $\Delta R^2$ | TabAutoPNPNet | 0.026 |
| | RMSE (%) | TabAutoPNPNet | 3.92 |
| | MAE (%) | TabFourierNet | 4.54 |
| Brazilian_houses | MAE (%) | TabAutoPNPNet | 8.67 |
| SGEMM_GPU_kernel_performance | $\Delta R^2$ | TabAutoPNPNet | 0.005 |
| | RMSE (%) | TabAutoPNPNet | 76.46 |
| | MAE (%) | TabAutoPNPNet | 81.66 |
| abalone | $\Delta R^2$ | TabChebyshevNet | 0.012 |
| | RMSE (%) | TabChebyshevNet | 0.74 |
| | MAE (%) | TabPNPNet | 1.25 |
| analcatdata_supreme | $\Delta R^2$ | TabAutoPNPNet | 0.007 |
| | RMSE (%) | TabAutoPNPNet | 12.93 |
| | MAE (%) | TabAutoPNPNet | 15.34 |
| diamonds | $\Delta R^2$ | TabAutoPNPNet | 0.002 |
| | RMSE (%) | TabAutoPNPNet | 1.78 |
| | MAE (%) | TabAutoPNPNet | 1.46 |
| house_sales | $\Delta R^2$ | TabAutoPNPNet | 0.015 |
| | RMSE (%) | TabAutoPNPNet | 5.38 |
| | MAE (%) | TabAutoPNPNet | 6.29 |
| medical_charges | $\Delta R^2$ | TabAutoPNPNet | 0.008 |
| | RMSE (%) | TabAutoPNPNet | 12.69 |
| | MAE (%) | TabAutoPNPNet | 19.65 |
| visualizing_soil | $\Delta R^2$ | TabPNPNet | 0.001 |
| | RMSE (%) | TabPNPNet | 3.19 |

Table 7.6: Performance improvements of proposed models over FT-Transformer for **numeric-only regression** tasks. Relative (%) improvements are reported for RMSE and MAE; absolute $\Delta R^2$ is shown for $R^2$.

| **Dataset** | **Metric** | **Best Model** | **Improvement** |
|---|---|---|---|
| Ailerons | $\Delta R^2$ | PNPNet | 1.926 |
| | RMSE (%) | PNPNet | 13.30 |
| | MAE (%) | PNPNet | 14.68 |
| Bike_Sharing_Demand | $\Delta R^2$ | AutoPNPNet | 0.010 |
| | RMSE (%) | AutoPNPNet | 1.06 |
| | MAE (%) | PNPNet | 2.54 |
| abalone | $\Delta R^2$ | PNPNet | 0.041 |
| | RMSE (%) | PNPNet | 2.56 |
| | MAE (%) | PNPNet | 2.50 |
| diamonds | $\Delta R^2$ | FourierNet | 0.001 |
| | RMSE (%) | FourierNet | 0.89 |
| | MAE (%) | FourierNet | 0.47 |
| elevators | $\Delta R^2$ | PNPNet | 0.022 |
| | RMSE (%) | PNPNet | 8.72 |
| | MAE (%) | PNPNet | 5.67 |
| house_16H | $\Delta R^2$ | PNPNet | 0.041 |
| | RMSE (%) | PNPNet | 1.73 |
| | MAE (%) | PNPNet | 1.62 |
| medical_charges | $\Delta R^2$ | ChebyshevNet | 0.011 |
| | RMSE (%) | ChebyshevNet | 15.59 |
| | MAE (%) | AutoPNPNet | 28.50 |
| pol | $\Delta R^2$ | AutoPNPNet | 0.0002 |
| sulfur | $\Delta R^2$ | PNPNet | 0.097 |
| | RMSE (%) | PNPNet | 15.13 |
| | MAE (%) | PNPNet | 8.87 |
| superconduct | $\Delta R^2$ | AutoPNPNet | 0.037 |
| | RMSE (%) | AutoPNPNet | 13.32 |
| | MAE (%) | AutoPNPNet | 14.71 |
| wine_quality | $\Delta R^2$ | FourierNet | 0.067 |
| | RMSE (%) | FourierNet | 1.82 |
| | MAE (%) | FourierNet | 3.84 |
| w | $\Delta R^2$ | FourierNet | 0.048 |
| | RMSE (%) | FourierNet | 0.76 |
| | MAE (%) | FourierNet | 2.95 |

### 7.8.2 Classification Tasks

Tables 7.7 and 7.8 present the performance improvements of the proposed models for the mixed-feature and numerical classification tasks, respectively. In classification tasks that combine categorical and continuous numerical features, the proposed models outperform FT-Transformer on 4 of 7 datasets. In classification tasks involving only continuous numerical features, the same models outperform the baseline on 9 out of 13 datasets.

Table 7.7: Performance improvements of proposed models over FT-Transformer for **mixed-feature classification** tasks. All improvements are reported as relative (%) changes with respect to FT-Transformer baseline scores.

| **Dataset** | **Metric** | **Best Model** | **Improvement (%)** |
|---|---|---|---|
| compas-two-years | Accuracy | TabChebyshevNet | 0.65 |
| | Precision | TabAutoPNPNet | 0.57 |
| | Recall | TabChebyshevNet | 2.30 |
| | F1 Score | TabChebyshevNet | 1.25 |
| default-of-credit-card-clients | Accuracy | TabChebyshevNet | 0.78 |
| | Precision | TabChebyshevNet | 0.75 |
| | Recall | TabAutoPNPNet | 4.49 |
| | F1 Score | TabChebyshevNet | 0.98 |
| electricity | Accuracy | TabAutoPNPNet | 3.09 |
| | Precision | TabAutoPNPNet | 2.74 |
| | Recall | TabAutoPNPNet | 3.64 |
| | F1 Score | TabAutoPNPNet | 3.20 |
| eye_movements | Accuracy | TabFourierNet | 3.47 |
| | Precision | TabFourierNet | 5.19 |

Table 7.8: Performance improvements of proposed models over FT-Transformer for **numeric-only classification** tasks. All improvements are reported as relative (%) changes with respect to FT-Transformer baseline scores.

| Dataset | Metric | Best Model | Improvement (%) |
|---|---|---|---|
| Diabetes130US | Accuracy | TabAutoPNPNet | 0.16 |
| | Precision | TabFourierNet | 0.07 |
| | Recall | TabPNPNet | 1.15 |
| | F1 Score | TabAutoPNPNet | 0.49 |
| MagicTelescope | Accuracy | TabAutoPNPNet | 1.20 |
| | Precision | TabChebyshevNet | 1.64 |
| | Recall | TabAutoPNPNet | 2.72 |
| | F1 Score | TabAutoPNPNet | 1.47 |
| bank-marketing | Accuracy | TabChebyshevNet | 0.37 |
| | Precision | TabChebyshevNet | 0.28 |
| | Recall | TabAutoPNPNet | 1.04 |
| | F1 Score | TabChebyshevNet | 0.39 |
| credit | Accuracy | TabAutoPNPNet | 3.08 |
| | Precision | TabPNPNet | 0.31 |
| | Recall | TabAutoPNPNet | 8.63 |
| | F1 Score | TabAutoPNPNet | 4.57 |
| default-of-credit-card-clients | Accuracy | TabChebyshevNet | 0.58 |
| | Precision | TabPNPNet | 0.68 |
| | Recall | TabAutoPNPNet | 5.04 |
| | F1 Score | TabFourierNet | 1.40 |
| electricity | Accuracy | TabAutoPNPNet | 2.84 |
| | Precision | TabFourierNet | 2.59 |
| | Recall | TabAutoPNPNet | 4.58 |
| | F1 Score | TabAutoPNPNet | 3.18 |
| eye_movements | Accuracy | TabFourierNet | 4.02 |
| | Precision | TabAutoPNPNet | 4.89 |
| | F1 Score | TabFourierNet | 2.30 |
| house_16H | Accuracy | TabChebyshevNet | 0.83 |
| | Precision | TabChebyshevNet | 0.52 |
| | Recall | TabChebyshevNet | 1.25 |
| | F1 Score | TabChebyshevNet | 0.88 |
| pol | Accuracy | TabChebyshevNet | 0.30 |
| | Precision | TabAutoPNPNet | 0.69 |
| | F1 Score | TabChebyshevNet | 0.29 |

### 7.8.3 Discussion

The benchmarking results provide strong empirical support for the central hypothesis of this work: *embedding explainable inductive structures within model design can simultaneously enhance predictive performance and explanatory coherence*. Across 52 datasets spanning regression and classification tasks, the proposed architectures achieved consist-

ent gains over the FT-Transformer baseline in 34 cases, demonstrating that periodic and non-periodic encoders capture complementary aspects of tabular structure. Importantly, these gains manifest across datasets with diverse feature compositions—numerical-only or mixed—confirming the robustness and adaptability of the approach.

Individually, **FourierNet** and **ChebyshevNet** outperform the FT-Transformer in numerous scenarios, validating the hypothesis that specialized encoders can disentangle distinct generative regularities within tabular data. The Fourier-based representation excels on datasets characterized by latent cycles or seasonality, whereas the Chebyshev-based encoder is effective for irregular, non-periodic dependencies. This division of representational labor embodies a form of *functional explainability*: each encoder exposes a distinct, explainable dimension of the model's reasoning that aligns with recognizable data phenomena.

When integrated, as in **PNPNet** and **AutoPNPNet**, the architectures achieve the most consistent overall improvements, confirming that fusing periodic and non-periodic representations enables models to reconcile structured and residual variability. These hybrid architectures enact the principle of *alignment* discussed earlier in the thesis—mediating between machine-learned abstraction and domain-grounded regularity. The modest but consistent advantage of AutoPNPNet over PNPNet further suggests that automatic feature-type inference can reveal subtle structures that manual feature partitioning may overlook, indicating that explainability is an emergent rather than prescribed property of the system.

From a computational standpoint, the inclusion of specialized encoders moderately increases the parameter count and training time. However, the trade-off between complexity and explainability remains favourable: the additional cost yields models that are not only more accurate but also more semantically transparent. The Fourier and Chebyshev encoders produce representations whose meaning is analytically tractable—namely, frequency components and polynomial transformations—thereby enabling a principled interpretation of the learned features. This directly addresses one of the persistent critiques of deep tabular learning: that predictive performance often comes at the expense of opacity.

Viewed through the broader lens of this thesis, these results demonstrate that explainability need not entail a loss in accuracy. On the contrary, architectures that internalize explainable structures—here, periodicity and smooth functional approximation—achieve superior generalization precisely because their inductive biases reflect the true organization of the data-generating processes. In this sense, performance improvement becomes an indicator of *faithfulness*: the model's internal reasoning aligns more closely with the domain's causal and temporal logic.

Overall, the findings suggest that **structural explainability**—embedding domain-coherent inductive priors into DL architectures—offers a productive path toward reconciling predictive power with intelligibility. The periodicity-based framework introduced here exemplifies how XAI principles can inform model design itself, transforming explainability from an external interpretive layer into a constitutive property of the system.

## 7.9 Explainability

Understanding how models arrive at their predictions is a foundational requirement for the responsible deployment of AI in industry and science. In domains where predictions inform high-stakes or mission-critical decisions, explainability is not a secondary concern but a condition for trust. Within the conceptual framework of this thesis, explainability

entails a form of *alignment*—a structured correspondence between the model's internal representations and the domain's intelligible regularities. The architectures introduced in this work—**FourierNet**, **ChebyshevNet**, **PNPNet**, and **AutoPNPNet**—are conceived precisely in this spirit: they aim to improve predictive accuracy while embedding forms of reasoning that are mathematically transparent and semantically grounded.

### 7.9.1 From Structural Encoding to Epistemic Transparency

By design, this work's models integrate specialized encoders that capture explainable patterns within tabular data. The Fourier and Chebyshev encodings do more than augment representational power—they *instantiate a logic of explanation*. Each encoding introduces an analytical structure that translates data variation into a domain-understandable form: frequency decomposition for periodic relationships and polynomial expansion for smooth non-linear dependencies. In this way, explainability becomes a property of the representational substrate itself, rather than a post-hoc interpretive layer.

Fourier encodings express the contribution of periodic features in terms of their constituent frequencies, allowing practitioners to trace model behaviour back to explainable temporal or cyclical patterns. Chebyshev encodings, by contrast, provide a flexible functional approximation that reveals the degrees of nonlinearity through which the model relates inputs to outputs. Both encoders preserve a direct correspondence to the original feature space: their transformations are analytically reversible and therefore *faithful* to the model's computational process. This structural faithfulness supports a more intelligible understanding of what the model has learned and why it behaves as it does.

### 7.9.2 Model-specific explainability Mechanisms

In **FourierNet**, each feature is represented by a set of learned frequency components. The magnitude of the learned weights associated with these components indicates which periodic patterns the model deems most relevant to the prediction. For instance, large weights on specific frequencies suggest that certain cyclical behaviors—such as daily, weekly, or seasonal—exert a strong influence on the outcome. Visualizing these spectral weight magnitudes thus transforms the abstract mathematics of frequency decomposition into a direct, domain-relevant explanation: the model predicts in part *because* a specific periodic rhythm dominates the data.

In **ChebyshevNet**, features are expanded into Chebyshev polynomial terms that capture varying degrees of nonlinearity. By examining the learned weights associated with each polynomial degree, I can identify the depth of curvature or interaction that drives the model's reasoning. A high weight on higher-degree terms indicates that the model relies on more complex, non-linear dynamics, while lower-degree terms capture smoother, near-linear relationships. This decomposition provides an explainable gradient from simplicity to complexity, aligning the mathematical expansion with the human conceptual continuum of linear versus non-linear dependence.

**PNPNet** merges these two interpretive pathways, producing a hybrid model in which periodic and non-periodic features are analyzed in separate branches before fusion. This separation enhances explainability by allowing feature importance to be assessed independently across structural types. Examining branch activations reveals whether periodic cycles or residual nonlinearities dominate the prediction. In this sense, the architecture itself becomes an epistemic instrument: its topology mirrors the decomposition of the

world into cyclical and aperiodic processes, thereby achieving a form of structural alignment between the model and the domain.

**AutoPNPNet** extends this paradigm by introducing an attention mechanism that automatically learns to weigh the relative contribution of each branch. The attention coefficients $\alpha_F$ and $\alpha_C$ quantify the model's reliance on periodic and non-periodic representations, respectively:

$$\boldsymbol{\alpha} = \text{softmax}\left(\mathbf{W}_{\text{att}}\left[\mathbf{h}_{F,L_F};\ \mathbf{h}_{C,L_C}\right] + \mathbf{b}_{\text{att}}\right), \tag{7.50}$$

where $\boldsymbol{\alpha} = [\alpha_F, \alpha_C]^\top$ and $\alpha_F + \alpha_C = 1$. A higher $\alpha_F$ indicates that the model attributes greater explanatory weight to cyclical regularities, while a higher $\alpha_C$ implies that non-linear interactions predominate. This mechanism operationalizes a meta-level form of explainability: the model *explains itself* by explicitly quantifying which internal reasoning pathway it prioritizes for a given prediction.

### 7.9.3 Gradient-Based and Visual Explanations

Beyond architectural explainability, this work's models also support gradient-based attribution and visualization of encoded components. The gradient of the output with respect to each input feature,

$$\frac{\partial \hat{y}}{\partial x_i} = \sum_k \frac{\partial \hat{y}}{\partial z_k} \cdot \frac{\partial z_k}{\partial x_i}, \tag{7.51}$$

can be decomposed across the encoded components $z_k$ (e.g., frequency or polynomial terms), revealing how each transformation contributes to the prediction. This enables practitioners to identify not only which features matter but *which aspect* of each feature—specific frequencies or polynomial degrees—drives the outcome.

Visualizing these gradients and weight magnitudes transforms abstract encodings into cognitively accessible explanations. In FourierNet, plotting the learned spectral weights highlights dominant cycles; in ChebyshevNet, the salience of polynomial terms reveals the degree of nonlinearity underlying specific predictions. These representations thus make the model's internal logic empirically inspectable and communicable to human experts.

### 7.9.4 Toward Structurally Grounded explainability

Taken together, the explainability mechanisms embedded in these models illustrate a central claim of this thesis: explainability and performance are not competing objectives but complementary consequences of epistemically structured design. By leveraging periodicity and polynomial expansion as intelligible mathematical priors, this work's models produce explanations that are both *faithful* to the underlying computation and *meaningful* within the domain. This is what I have termed *structurally grounded explainability*—a form of explainability that arises from the alignment between representational geometry and the organization of the world it models.

In practical terms, this alignment enables stakeholders to engage with model reasoning at multiple levels of abstraction: from frequency components and polynomial terms to aggregated attention weights and global sensitivity analyses. Conceptually, it demonstrates that explanations can be designed into the architecture itself, transforming them from retrospective rationalizations into constitutive properties of intelligent systems. In doing so, the Fourier–Chebyshev family of models advances the broader thesis that explainability

is not a cosmetic layer added to black-box systems, but a design principle that redefines what it means for an AI model to *understand* and to be *understood*.

## 7.10 Limitations

This study acknowledges certain limitations. The performance of the specialized encoders may depend on hyperparameter choices such as the number of frequency components and the degree of Chebyshev polynomials, indicating hyperparameter sensitivity. For extremely large datasets, the increase in the number of parameters may lead to longer training times and higher memory requirements, posing scalability challenges. While this work's models handle categorical features using embeddings, the specialized encoders are primarily designed for numerical data. Extending their capabilities to better integrate categorical features remains an area for future work.

Potential directions for future research include developing methods to learn the optimal number of frequency components and polynomial degrees during training, thereby adapting encoder parameters. Extending specialized encoders to process categorical features, possibly through categorical-specific transformations, would more effectively enhance the integration of categorical features. Exploring techniques to reduce computational overhead, such as pruning or quantization, without significantly degrading performance, could enable model compression. Combining this work's approach with other advanced models, such as attention mechanisms or graph neural networks, to further improve performance on tabular data represents a potential for hybrid architectures.

## 7.11 Conclusions

This work presented a family of neural architectures—**FourierNet**, **ChebyshevNet**, **PNPNet**, and **AutoPNPNet**—designed to enhance tabular data modelling by embedding explainable structural priors directly into the learning process. Each model embodies a distinct yet complementary principle of structured representation: FourierNet captures *periodic regularities* through frequency-based encodings, while ChebyshevNet models *non-linear dependencies* via polynomial expansions. Their integration into PNPNet enables simultaneous reasoning over both periodic and non-periodic structures, and AutoPNPNet extends this capability by employing an attention mechanism that adaptively balances these components without requiring manual feature separation.

Extensive benchmarking across 53 datasets demonstrates that these architectures outperform the current SotA DL model, FT-Transformer, on 34 tasks spanning regression and classification. These results confirm the hypothesis that *structurally grounded inductive biases*—when aligned with the statistical and temporal logic of real-world data—enhance not only predictive accuracy but also the interpretive transparency of the learned representations. The specialized encoders yield models that are sensitive to the internal organization of data while remaining intelligible to human reasoning, thereby addressing a key limitation of traditional deep tabular methods.

Beyond performance, the explainability of these models represents their most significant contribution. By decomposing features into analytically meaningful bases—such as Fourier frequencies and Chebyshev polynomials—the proposed architectures make the learned representations directly explainable. The weight magnitudes associated with frequency components or polynomial degrees reveal which periodic or non-linear patterns

drive predictions, while the attention mechanism in AutoPNPNet exposes the model's allocation of reasoning between cyclical and residual dynamics. These mechanisms translate otherwise opaque internal computations into *structurally faithful explanations*, aligning the model's reasoning with domain-understandable concepts such as cycles, trends, and deviations.

From the broader perspective of this thesis, these findings substantiate the central claim that explainability can be achieved not merely through post hoc interpretation but through *design*. Embedding domain-coherent structures—here, periodicity and smooth functional approximation—within model architecture transforms explainability from an accessory into a constitutive property of intelligent systems. The Fourier–Chebyshev framework thus exemplifies how the threefold requirement of *faithfulness, intelligibility, and alignment* can be operationalized in practice.

Ultimately, this study underscores the industrial significance of structurally aware explainability. By grounding neural reasoning in recognizable temporal and functional patterns, these architectures provide not only accurate but also auditable models—suitable for deployment in domains where explainability is inseparable from reliability. In this sense, the proposed models bridge the gap between theoretical principles and applied practice, demonstrating that explainable design is both an epistemic commitment and a pragmatic advantage. The chapter thereby extends the thesis's overarching argument: that responsible, high-performance AI must evolve toward systems whose reasoning is as transparent as it is effective.

# III

# Reflections Moving Forward

# 8

# The Principle Of Appropriate Model Complexity

***This chapter is based on:*** *M. Rizzo et al. 'Stop overkilling simple tasks with black-box models, use more transparent models instead'. In:* Pattern Recognition and Artificial Intelligence*. Ed. by C. Wallraven, C.-L. Liu and A. Ross. Singapore: Springer, 2025, pp. 279–293. ISBN: 978-9819787012. DOI: 10.1007/978-981-97-8702-9_19. URL: https://doi.org/10.1007/978-981-97-8702-9_19*

As this thesis transitions from specific analyses to a concluding synthesis, its central argument converges on a fundamental tension in modern AI. The inquiry began with a theoretical question — *what constitutes an explanation?* — and unfolded through a progressive reconciliation of theory and practice. A consistent pattern emerged: explainability is not an accessory to model performance but a constitutive dimension of model design. This insight forces a confrontation with the implicit bargain at the heart of ML: the trade-off between a model's predictive power and our capacity to comprehend its reasoning.

This tension reverberates across the three research frontiers identified in this thesis — *intelligibility*, *alignment*, and *faithfulness* — each revealing a distinct role of explanations and, consequently, a distinct notion of what makes an AI model *appropriate*. In **medicine**, where the pursuit of intelligibility frames explanation as an *act of communication*, models must translate machine reasoning into a language that sustains dialogue between algorithmic and clinical judgment. Excessive opacity fractures this communicative chain and undermines trust. In **security**, particularly in the context of LLMs, the focus shifts to alignment: explanation becomes an *act of knowledge construction*. Explanations are

valuable not for their rhetorical persuasiveness but for their role in scaffolding reasoning — guiding both human and machine agents toward actionable insights. The architecture of LLMs thus embodies both promise and peril: their emergent capacities inspire optimism, yet their internal opacity complicates verification. Here, complexity is justified only to the extent that it supports structured reasoning and accountability. Finally, in **industry**, faithfulness reframes explanation as an *act of design*. Models must be auditable and compliant with governance constraints; explainability becomes a structural feature, and complexity a design variable constrained by the dual requirement of transparency and integration.

Together, these lessons converge in a general synthesis: the epistemic value of a model is proportional not to its magnitude, but to the *appropriateness* of its complexity. This yields an operational principle for trustworthy AI: the *Principle of Appropriate Model Complexity (PAMC)*.

## 8.1 Formal Definition

The cumulative analyses of this thesis point to a central insight: the pursuit of explainability is, at its core, a matter of *complexity governance*. Modern AI has often taken the opposite path, equating progress with scale—treating complexity as a virtue in itself. The advent of LLMs epitomizes this trend. Their immense parameterization and emergent linguistic coherence have produced a new paradox: systems that appear more intelligible than ever yet remain more opaque than any model that has come before. The PAMC is introduced precisely to redress this imbalance, formalizing the intuition that scale must be epistemically warranted.

**Definition 8.1** (Principle of Appropriate Model Complexity)**.** Let a model $M$ approximate or infer a target function $f$ over domain $\mathcal{D}$, evaluated by two classes of criteria:

- a **performative requirement** $\mathcal{P}(M) \geq P^*$, ensuring adequate predictive or functional competence;
- an **epistemic requirement** $\mathcal{E}(M) \geq E^*$, ensuring that the model's operations remain intelligible, faithful, and aligned to human understanding.

Then the model's *appropriate complexity* $\mathcal{C}^*$ is the minimal complexity $\mathcal{C}(M)$ satisfying both thresholds:

$$M^* = \arg\min_{M \in \mathcal{M}} \mathcal{C}(M) \quad \text{s.t.} \quad \mathcal{P}(M) \geq P^* \ \text{ and } \ \mathcal{E}(M) \geq E^*.$$

A model is *epistemically optimal* when no simpler model meets both the performance and epistemic constraints of its domain.

## 8.2 Usefulness and Implications of the Principle

The PAMC treats model complexity as a measurable and optimizable dimension of design. The core idea is that models should be only as complex as needed to achieve faithful, intelligible, and aligned behavior in their intended use context. This reframing has three practical implications: complexity must be *measured*, *modulated*, and *justified*.

### 8.2.1 Operationalizing Complexity

To apply the PAMC, complexity must first be rendered observable. In practice, model complexity can be characterized along three complementary axes: *structural*, *functional*,

and *epistemic*. *Structural complexity* refers to the formal properties of the model — such as the number of parameters, depth of layers, connectivity, or nonlinearity of the mapping between inputs and outputs. These properties can be quantified directly (e.g., via model capacity measures such as VC dimension, Lipschitz constants, or sparsity ratios) or indirectly via compression tests and information-theoretic proxies that estimate the model's effective dimensionality. *Functional complexity* captures the amount of reasoning or representational capacity actually engaged during inference. It depends on the task configuration rather than on the architecture alone. In the case of LLMs, for instance, functional complexity can be modulated by the length and compositionality of prompts, the scope of retrieval, or the depth of reasoning. This form of complexity can be quantified by tracking activation sparsity, context entropy, or token-level attention dispersion across tasks. Finally, *epistemic complexity* measures how difficult it is for a human — or another model — to form reliable expectations about the system's behavior. It can be estimated through the stability and faithfulness of explanations: models that exhibit highly unstable attribution maps or divergent causal graphs across small perturbations are epistemically more complex. Recent approaches quantify this using explanation-consistency metrics, counterfactual-stability indices, or concept-alignment measures that compare latent representations to domain ontologies. Together, these measures allow complexity to be treated as a variable in the design process rather than an uncontrollable byproduct of scaling.

### 8.2.2 Practical Implications of Choosing Lower Complexity

Selecting a less complex model entails both benefits and trade-offs that can be systematically analyzed under the PAMC. On the positive side, models with lower structural and epistemic complexity are typically more transparent, data-efficient, and stable under distributional shifts. They reduce the cognitive and institutional overhead associated with validation, auditing, and regulatory compliance. They also facilitate human oversight, as domain experts can more easily detect spurious correlations or biases.

However, lower complexity may reduce representational flexibility and limit performance on high-dimensional or weakly structured problems. The PAMC does not deny this trade-off but requires that any increase in complexity be epistemically warranted — that is, justified by measurable improvements in explanatory coherence, stability, or alignment with domain structure. This transforms complexity from a mere engineering decision into a question of epistemic proportionality. Rather than optimizing solely for predictive accuracy, practitioners are encouraged to evaluate whether additional layers of abstraction genuinely improve the explainability or robustness of the model's reasoning process. The principle thus supports a form of *complexity budgeting*: allocating expressive capacity where it yields epistemic returns, and constraining it where it merely increases opacity. For instance, an LLM-based information retrieval pipeline can remain tractable by externalizing certain reasoning steps to symbolic modules or verifiable retrieval components. The overall system retains expressive power while maintaining a transparent and auditable inferential structure.

### 8.2.3 Embedding the Principle in Practice

Applying the PAMC throughout the AI lifecycle requires explicit mechanisms for calibration and justification. During **model design**, architectural decisions should be guided by the structure of the task: sparse, monotonic, or graph-constrained architectures are

appropriate where relationships are well-understood, whereas deep or transformer-based structures are justified only when high-dimensional correlations are essential.

In **training**, regularizers and inductive constraints can be used to penalize unnecessary complexity—e.g., feature redundancy, high attention entropy, or inconsistent causal attributions. In **evaluation**, performance metrics should be complemented with complexity–faithfulness trade-off analyses. Such analyses quantify whether marginal performance improvements are accompanied by losses in intelligibility or stability. For example, validation reports can include an "explanatory efficiency curve" that shows accuracy gains as a function of complexity measures, such as parameter count, explanation variance, or activation entropy. This provides an empirical basis for deciding when additional capacity ceases to yield real benefit. At the **deployment and governance** level, the PAMC motivates the inclusion of a "complexity justification statement" alongside conventional model documentation. This report specifies why the chosen level of expressivity is epistemically appropriate for the application's risk profile, how it was measured, and which controls ensure explainability. In regulated domains—such as healthcare, finance, or critical infrastructure—such documentation could function as an auditable artefact, linking model architecture to oversight procedures. Finally, in **human–AI interaction**, the appropriate level of complexity determines how models communicate their reasoning. When users must make time-critical or high-stakes decisions, explanations should compress internal complexity into domain-relevant abstractions, preserving causal faithfulness while remaining cognitively accessible. Surrogate models, structured rationales, or interactive visualizations can translate the behavior of complex systems into forms compatible with expert judgment.

### 8.2.4 Modulating the Complexity of Emergent Abilities

Models with *emergent abilities*, such as LLMs, pose specific challenges for applying the PAMC. In these systems, reducing architectural complexity typically reduces capability, and emergent behaviors introduce additional opacity. Consequently, control shifts from structural complexity—such as the number of parameters—to *functional complexity*, which describes how much of the model's latent capacity is activated during inference. Functional complexity can be adjusted by designing prompts that specify their scope, abstraction level, and compositional structure. It can also be influenced by the size of the context window, the use of external memory or retrieval mechanisms, and the configuration of inference parameters such as reasoning depth or temperature. The integration of structured pipelines or external tools can also regulate functional complexity by decomposing tasks into explainable sub-components.

Even when the model's internal operations remain opaque, intelligibility can be improved by embedding the model within structured frameworks such as retrieval-augmented generation, symbolic reasoning layers, or modular tool integration. These scaffolds do not simplify the model itself but make its functional behavior more explainable. Determining the appropriate level of functional complexity requires empirical calibration, which involves identifying the point at which additional capacity no longer yields significant gain relative to the costs of explainability. This calibration can be evaluated by tracking improvements in explanation quality or user comprehension as a function of model configuration, measuring the stability of outputs under small perturbations to input or prompt structure, and verifying whether latent representations correspond to meaningful domain structures or ontologies.

## 8.3 Epistemic Governance and the Complexity Frontier

Ultimately, the PAMC reframes AI development as a problem of *epistemic governance*. The relevant question is no longer "*how complex can our models become?*" but "*how much complexity is epistemically and ethically justified*, given human interpretive capacities and the moral weight of the decisions involved?"

In this light, emergence and intelligibility must co-evolve. As models acquire new representational and generative powers, new scaffolds of understanding—conceptual, methodological, and institutional—must emerge in tandem. Appropriate complexity thus marks not a static limit but a dynamic equilibrium between the growth of model capacity and the growth of human comprehension. At this equilibrium, explainability ceases to constrain intelligence and becomes its precondition instead.

The remainder of this chapter empirically substantiates the PAMC, showing that an appropriate governance of complexity can yield operative models simultaneously accurate and explainable. In doing so, it outlines a route toward an *epistemically sustainable* AI—one in which complexity serves understanding rather than overwhelming it. The PAMC thus offers a path to reconcile the promise of FMs with the demands of responsible knowledge systems. In an era dominated by LLMs, sustainability depends not on further scaling, but on scaling *appropriately*: designing architectures whose complexity scales in proportion to the depth of understanding they enable.

## 8.4 An Invite to Stop Overkilling Simple Tasks With Complex Black-box Models

The present study advances a complementary approach grounded in the PAMC: rather than retrofitting explanations onto highly complex black-box models, I advocate for designing models whose structure is inherently more explainable, without sacrificing predictive performance. By leveraging simpler architectures that encode domain knowledge and salient task features, it is possible to achieve models that are both accurate and transparent. This work illustrates this approach through a practical, industry-relevant example, demonstrating that judiciously selecting model complexity enhances explainability while mitigating the hidden costs of unnecessarily opaque systems.

### 8.4.1 Task and Approach

To illustrate the design strategy of this work, I analyze a straightforward real-world scenario and examine how the concepts of accuracy and explainability discussed above affect it. The target task of this study is to classify the ripeness of banana crates on a scale from 1 (least ripe) to 4 (ripest) (see Fig. 8.1 for an example).

In the proposed approach, I design for competitive accuracy and explainability simultaneously. To tackle the classification task, I select a pool of three DL methods: (i) a simple Convolutional Neural Network (CNN) model with three convolutional blocks, (ii) a pre-trained convolutional model based on the MobileNetV2 framework [161], and (iii) a pre-trained Vision Transformer (ViT) [42]. As I will show, the latter yields nearly perfect results and is the best neural model among the proposed methods in this work; however, none of the XAI methods I have tried can accurately explain its predictions. On the other hand, I demonstrate that the proposed approach, based on simple color features


Figure 8.1: Ripeness stages for crates of bananas from least ripe (1) to ripest (4).

and a fine-tuned Decision Tree (DT), can achieve competitive accuracy while exposing the information needed (i.e., the *evidence*) to produce adequate and global explanations.

### 8.4.2 Contributions

The experiments conducted in this work reveal that three neural models quickly achieve high accuracy, raising questions about the necessity of complex, opaque methods for tasks that appear to be simple. This finding suggests the viability of simpler, more comprehensible models that do not compromise accuracy. Key contributions of this work include:

- Establishing design principles for ML problems, prioritizing explainability;
- Analyzing DL methods for an explicative classification task, focusing on accuracy and explainability with a selection of models providing a broad view of the task;
- Demonstrating the effectiveness of a simpler, more transparent DT model with minimal feature engineering in solving the same task;
- Conducting a user study to verify explanations align with stakeholders' needs;
- Releasing code and self-collected dataset[1] to support reproducibility and extension of the research.

## 8.5 Related Work

This research intersects two significant areas: (i) enhancing explainability in AI and (ii) optimizing fruit ripeness grading.

**Explainable AI.**

DL models are often criticized for their opacity, which makes explaining their predictions challenging. Efforts to derive explanations have employed various techniques, including gradient information [165], attention scores [10], and model-agnostic methods such as LIME and SHAP [147, 112]. However, these explanations have faced reliability challenges in several contexts. Alvarez-Melis et al. [5] report that SHAP and other model-agnostic saliency-based methods are susceptible to instability, leading to significant saliency differences in the face of minor input modifications that do not alter the overall prediction. Adebayo et al. [4] demonstrate that some gradient-based methods tend to

[1] https://github.com/matteo-rizzo/explainable-banana-ripeness-classification

behave like edge detectors, generating dangerously misleading visual maps, and reveal that these can be manipulated in unexpected ways. Other authors have highlighted the unreliability of several gradient-based methods, as the dependence on reference points to determine saliency makes the heatmaps highly sensitive to minor input transformations [95]. Additionally, substantial shifts in key features can lead to zero gradients due to the models' nonlinearity [130]. The authors of [166, 83] criticize the use of attention weights to determine importance, discovering that multiple attention distributions yield identical results. They also find a weak correlation between attention weights and other saliency measures, thereby questioning the reliability of this approach. Despite these issues, which may affect these XAI methods to varying degrees depending on the specific problem, I believe the primary limitation of these techniques is their inability to provide a semantic explanation of the model's decision that effectively conveys a clear, unambiguous meaning to the human stakeholder. Referring to the nomenclature introduced in the theoretical framework (Chapter 3), they could be considered as *evidence* extractors, but more research and effort are needed to bridge the gap between these extractors and faithful interpretations, ultimately necessary to deliver true explanations. Despite these issues, SHAP remains a leading approach to saliency-based explainability, offering both local and global model explanations. Because of this, in this study, I compare the use of SHAP to explain DL models with the more transparent approach adopted here.

Rudin et al. [159] emphasize the importance of using transparent models for high-stakes tasks. In line with this, the research presented here advocates simpler, transparent models that still achieve satisfactory performance, along with a strategy for faithful explanation.

**Fruit Ripeness Recognition.**

Grading fruit ripeness has been addressed using statistical methods [120], traditional ML [197], and DL approaches [162]. DL methods, known for their high accuracy and minimal feature engineering requirements, are currently among the top-performing methods. Rizzo et al. [152] provide a comprehensive survey on fruit ripeness grading. Recent trends focus on minor accuracy improvements, often overlooking the importance of explainability.

## 8.6 Designing for Explainability

The proposed guidelines aim to identify the problem features most intuitive to stakeholders and process them with minimal intervention using the simplest ML method adequate for the task. "Simplicity", in this case, relates to the number of parameters regulating the model (the lower, the better) and its reliance on human-understandable processing of the features (the more, the better).

In particular, I want to build a pipeline from raw data to prediction, with each step as transparent as possible. The proposed design process follows these high-level steps: (i) understand the task to be solved by the ML method, the available data, and the stakeholders of the final product; (ii) for each stakeholder, discuss which attributes they consider relevant in solving the task and define which features can be considered part of an explanation; (iii) find an ML model that is powerful enough to process the features but also offers the possibility to extract interesting evidence with a reasonable effort. The evidence must suggest an interpretation that is faithful by design to how the model works and

possibly aligns with human intuition for plausibility ; (iv) test model performance and effectiveness of the generated explanations: the model should provide competitive accuracy with the state-of-the-art, while also satisfying the expectations of the stakeholders with the produced explanations. I find that a user study is an effective way to obtain qualitative evidence of the proposed XUI's efficacy.

Step (ii) is perhaps the most challenging point, especially when very little problem-specific knowledge is available to the stakeholders. In this scenario, a preliminary analysis of the performance of top black-box models can indicate the task's difficulty. If the specific task reveals intuitive features that can be leveraged to solve it, a model that tends toward transparency is worth considering. Intuitiveness is critical to optimizing the design and to reaching a final explanation that is faithful to the model's behavior and plausible to the human stakeholder. On the other hand, I acknowledge that finding meaningful features or even just effective data representation can be challenging for some tasks. In Natural Language Processing, for example, handcrafting general, context-sensitive, and human-understandable features is often very difficult or impractical, partly due to the inherent complexity of natural languages. That is why I advocate reasoning about an ML problem and trying a broader explainability-driven approach, especially when the task is simple. For some tasks, simple or explainable solutions may not yet exist. The following sections demonstrate how I applied these guidelines to the example task in this study.

### 8.6.1 Task Definition, Stakeholders, and Data

From a practical perspective, this work deals with a multiclass image classification task. The stakeholders of this work are workers at the wholesale fruit market of Treviso, Italy, who are interested in automating the ripeness grading of banana bunches. Currently, operators manually label bunches according to ripeness (1 to 4, least to most ripe; see Fig. 8.1). All the bananas within a crate are assumed to be in the same ripeness stage. The ML classifier developed in this work would assist operators in labeling large numbers of incoming crates. Moreover, this is the first step toward digitalizing the fruit processing pipeline, from fruit quality inspection and assessment to online sales. Given the assessment step's impact on fruit pricing, stakeholders emphasized the importance of maintaining transparency in the grading process to enable human oversight.

I collected an *ad hoc* dataset comprising 927 images to develop the ML solution, with a reasonable balance across the four ripeness classes. The dataset was manually labeled by operators who perform quality assessments of incoming products. To understand human performance on this classification task, I also asked three operators to re-label a subset of images from the dataset. More technical details on the data are provided in Section 8.7.1, while human performance is reported in Section 8.8.

### 8.6.2 Feature Selection

After consultation with stakeholders, I determined that color is the most reliable and intuitive indicator of banana bunch ripeness. Images are encoded in the RGB color space, a well-known color model with solid theoretical foundations in human color perception. Since color is the most informative feature in this dataset, I process the images to extract color information and train a classifier to classify ripeness stages using these features. Section 8.7.1 details how this information is extracted and used in the proposed solution.

### 8.6.3 On the Choice of Models

I select state-of-the-art DL-based methods and simpler, more transparent classifiers for this task. Testing DL models provides an estimate of the best achievable performance and an indication of the problem's difficulty. As stated, the objective of this work is to select the model of the lowest complexity that achieves adequate performance while preserving as much transparency as possible. I selected a DT, a Support Vector Machine (SVM) with different kernels, and a multinomial Naive Bayes (NB) classifier as baseline models for comparison, ultimately choosing the DT as the best-performing model.

I point out that the DT learns discriminative rules that partition the feature space into sub-spaces corresponding to each target class (*i.e.*, the ripeness stage). By extracting color information in the RGB space and limiting the number of features, I can obtain a *global* explanation that maps each ripeness stage to specific regions of the color space. I emphasize that this explanation is *faithful*, in the sense that it accurately captures the DT's reasoning process, and *plausible*, in the sense that it aligns with human understanding of the problem. These characteristics make this strategy effective concerning point (iii) in this work's guidelines.

### 8.6.4 Testing for Accuracy and Explainability

I compare the performance of the baseline models (DT, SVM, NB) and select DT as the best compromise between the complexity and the intuitiveness of the explanation that can be derived from it, as discussed in the previous section. The NB classifier performs worse than the DT. Conversely, the SVM with a high-degree polynomial kernel achieved slightly better results (less than 0.5% improvement in accuracy and F1-score). However, given the minimal difference in results and the complexity of the SVM's decision boundaries, the DT appears to be a better choice. The complete results of these tests are reported in the supplementary material. Additionally, I compare the DT with state-of-the-art DL models that are well-established off-the-shelf solutions for this task, in line with recent trends in Computer Vision (CV). Results are reported in Section 8.8, showcasing that the DT achieves competitive performance and is well above human classification performance. Ultimately, I aim to evaluate the effectiveness of the generated explanations for the stakeholders of this work. To achieve this, I conducted a user study to investigate users' preferences regarding the generated explanations. More details on the results are provided in Section 8.8.3 while the complete questionnaire is reported in the supplementary materials.

## 8.7 Methods and Explanations

### 8.7.1 Data Processing

The dataset used in this study comprises 927 RGB images of banana crates, captured at 4160 x 3120 pixels with a CZUR Shine Ultra scanner under consistent lighting conditions. The images are balanced across classes and were resized to 224 x 224 pixels for compatibility with pre-trained models and to facilitate reasonable inference times. The dataset was augmented with random transformations, including rotation, affine and elastic morphological transformations, cropping, Gaussian blur, and perspective changes, to mimic real-world scenarios such as smartphone photography. About half of the dataset was augmented and incorporated into the training set; further details are in the supplementary ma-

terial. Visual inspection of the dataset revealed noise, particularly along the boundaries of the crates. To address this, I applied semantic segmentation using the SLIC algorithm [2], thereby improving results across all methods. This work's report focuses on the outcomes with segmented images. For the selected DL models, feature extraction is automated from the raw RGB input. However, I employed minimal feature engineering for the DT, focusing on color features. Each image was represented by the normalized average color values of its R, G, and B channels. I also adjusted the luminance by converting the images to YUV space, normalizing the Y channel, and then reverting to RGB. This process is important because RGB embeds luminance in its channels, whereas YUV uses a separate luminance channel.

### 8.7.2 Deep Learning Approach

To address banana ripeness classification, I run and compare three neural approaches. The first architecture consists of a simple CNN with three convolutional blocks, each comprising two 2D convolutions and max-pooling, interleaved with ReLU activations. The convolutional layers extract features, which are fed to a three-layer feed-forward ANN that outputs the final prediction. Before being processed by the CNN, the data is normalized to the mean and standard deviation. The second architecture I consider is the pre-trained MobileNetV2 network [161]. Still convolutional by nature, the strategy at the core of this method is based on depth-wise convolutions and inverted residual connections. The designers aimed to build a powerful, pretrainable model for low-tier devices. The third architecture I examine is the Vision Transformer (ViT) [42]. Transformers [181] are neural architectures based on multi-head attention [10], widely studied and employed by the NLP community [116]. This architecture has recently been applied to CV tasks with various strategies (see [107] for a survey). Briefly, ViT splits images into fixed-size patches and linearly embeds them. Positional embeddings are then added to retain positional information, after which the resulting sequence of vectors is fed to a standard Transformer encoder. Classification is achieved by adding a learnable "classification token" to the sequence. In the presented experiments, I use the `vit-base-patch16` model [195], which was pretrained on ImageNet.

**Deep Learning Explainability Strategy.**

As previously mentioned, I used SHAP [112] to explain the DL models' predictions. When working with images, SHAP can generate heat maps (which constitute the XUI) to provide explanations to the user. These are supposed to describe the importance of each pixel in the image toward the model's prediction. Intuitively, warm colors indicate the regions of the image that contributed the most to the prediction. In contrast, colder colors indicate areas that contributed negatively to the prediction of the same class. Example explanations generated with SHAP are presented in Fig. 8.2. As mentioned in Section 8.5, feature importance heatmaps, commonly generated with saliency methods, can lead to potentially different (and deceptive) interpretations by end-users that may not accurately reflect the model's actual decision-making process. This is an alarming condition in which the explanations convey to the user a "convincing lie" about the model's behavior. The following sections show how the design of this work addresses faithfulness and plausibility.

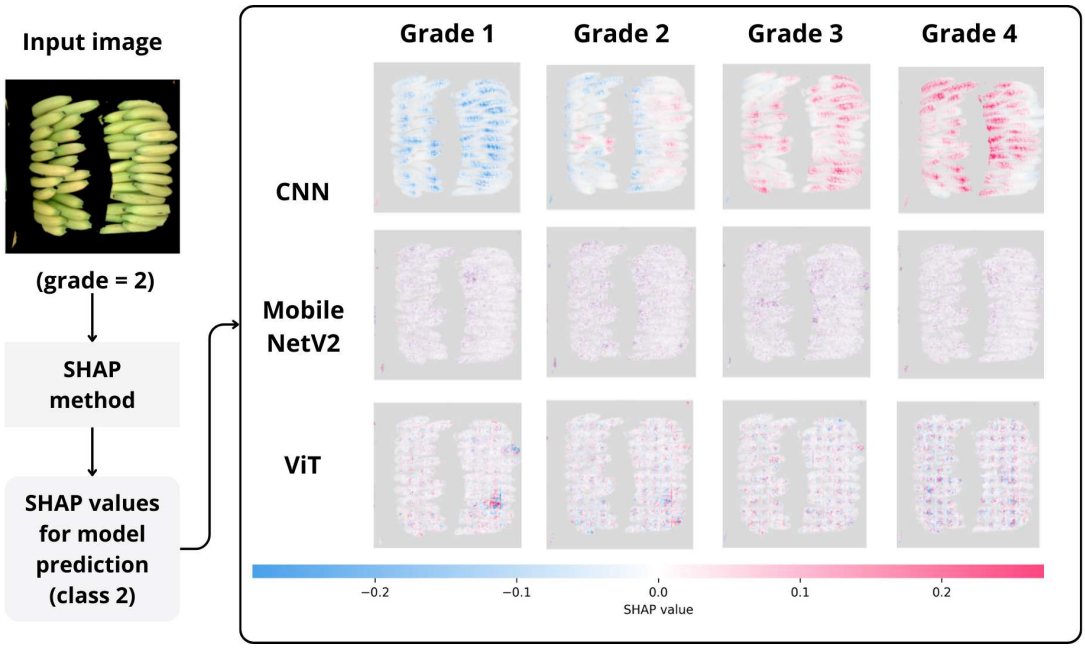



Figure 8.2: Examples of explanations for DL models generated using SHAP.

### 8.7.3 Decision Tree

In contrast to the examined DL methods' internal complexity, I propose tackling the same task using a simple, more transparent DT-based classifier. In particular, I use the scikit-learn implementation of the CART algorithm [15].

**Explainability Strategy.**

One may argue that a DT is an intrinsically explainable model. I argue that there is no such thing as intrinsic explainability: a transparent model still needs to provide some explanation that is somewhat understandable to users and answers their 'why' questions. As stated in Chapter 3, models provide evidence, and generating a good explanation requires giving that evidence semantic meaning, a process called *interpretation*. Moreover, different end users likely have distinct requirements for explainability. An *explanation interface* must then convey the interpreted information to the users. For example, ML experts may be satisfied with understanding the range of feature values mapped to each target class (in this work's case, the RGB values). Non-expert users may need these rules further processed to be presented more clearly. Serving explainability is intuitively easier for specific models, such as those with a few parameters, though this has not yet been formalized in the literature. Admittedly, a DT has a very intuitive and faithful interpretation: for every non-leaf node, the DT learns a threshold value for one of its given features, thus producing two children (above and below the threshold). In the present case, each instance is classified by following a path to a leaf labeled with a specific ripeness value. Conveniently, the set of rules along the traversed path defines an area within the RGB color space, which is part of the explanation in this work. Binding the explanation to the intuitive process of discriminating between banana crates by color (as the stakeholders in this work do) lays the groundwork for plausibility. Albeit simple for relatively shallow trees, the decision paths can grow exponentially for features with complex interactions. As anticipated, such numerical features split within the DT can still appear opaque to the average user. Thus, I

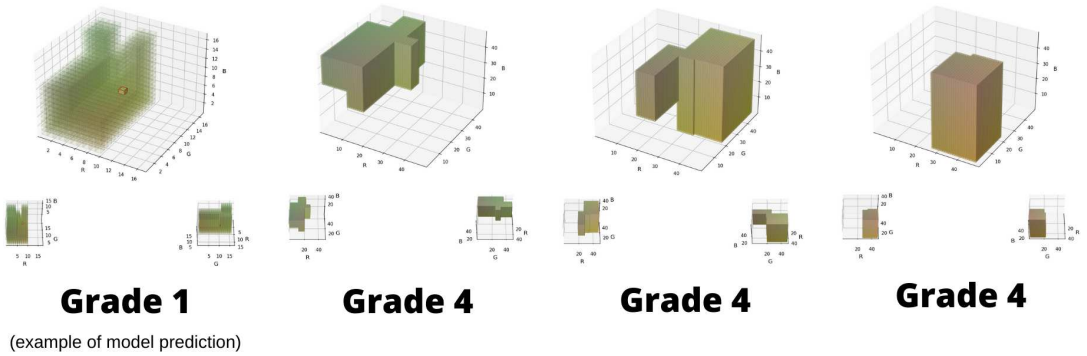



Figure 8.3: Explanation generated from the constraints imposed by the DT on the RGB color gamut. The four grades correspond to distinct areas within the gamut.

| | *Accuracy* | *Precision* | *Recall* | $F_1$ |
|---|---|---|---|---|
| Decision Tree | 0.9716 (± .0104) | 0.9723 (± .0106) | 0.9678 (± .0119) | 0.9697 (± .0110) |
| CNN | 0.9349 (± .0115) | 0.9298 (± .0131) | 0.9308 (± .0123) | 0.9377 (± .0123) |
| MobileNet V2 | 0.9743 (± .0046) | 0.9726 (± .0046) | 0.9717 (± .0054) | 0.9718 (± .0049) |
| ViT | 0.9967 (± .0015) | 0.9960 (± .0020) | 0.9966 (± .0017) | 0.9962 (± .0018) |
| Human Performance | 0.7500 (± .0589) | 0.7588 (± .0453) | 0.7500 (± .0589) | 0.7519 (± .0524) |

Table 8.1: Macro-averaged performance metrics for the models averaged over ten random seeds (standard deviation in brackets).

further explain this work by devising an XUI that is human-understandable and tested accordingly. More specifically, I use the rules extracted from the decision path as constraints on the RGB gamut to identify regions of RGB space that correspond to the four ripeness classes. Hence, it is straightforward to describe each unknown input data point by its average color in the 3D RGB color space and determine to which region it belongs. This plot is the proposed explanation for the DT's behavior. Fig. 8.3 is an example visualization of the whole process (more examples are reported in the supplementary material). It is worth stressing that the area of the color space extracted from the decision rules learned by the DT is, by definition, a *global* explanation. Thus, this strategy allows us to identify which colors are unequivocally associated with each label class. One benefit of such an explanation is the ability to validate the classifier's behavior. Unexpected colors would appear in the proposed XUI, pointing out a negative bias in the model.

## 8.8 Experiments

In this section, I compare the performance achieved by the methods employed in this work. First, I analyze the classification metrics achieved by the three DL-based models and the DT. Then, I study the explanations generated according to the strategies proposed in Sections 8.7.2 and 8.7.3, and compare them through a user study involving stakeholders for the task of banana ripeness classification in a real fruit market.

### 8.8.1 Performance

To assess the ability of the selected models in this work to produce correct predictions, I use commonly used classification metrics: accuracy, macro-averaged precision, macro-averaged recall, and macro-averaged F1 score. All methods are tested using 5-fold cross-validation, repeated ten times with different random seeds to enhance the robustness of the results. Table 8.1 presents the results obtained with both deep learning methods and the DT. I also report human performance, defined as the average score across three stakeholders in classifying a balanced dataset of 300 randomly sampled images from the original non-augmented dataset ($\sim 20\%$). It is evident that all methods achieve excellent results, with all metrics surpassing the 90th percentile scores and outperforming a human baseline. It is worth remembering that these results are achieved on datasets augmented with images that have undergone various transformations, which makes them more robust at the cost of small performance decreases.

Further detailed in the supplementary material, the error analysis reveals that errors always occur because the classifiers select an adjacent class (*e.g.*, class 2 instead of class 1). The ViT model achieves near-perfect scores across all metrics among the selected methods. The DT also obtained outstanding results, though this required more effort (including standardizing luminance and conducting an extensive grid search). Nevertheless, this process allows the DT to have results comparable to those of MobileNetV2.

### 8.8.2 Explainability

I compare the SHAP explanations for the DL models with the handcrafted RGB-based explanations designed for the DT. Fig. 8.2 and 8.3 compare the two types of explanations for the same input. It is clear that the masks produced by SHAP do not highlight meaningful features of the image. Indeed, the highlighted regions appear random. Not only that, in this work's case, SHAP's visualization for the CNN consistently showed the same result across all classes, seemingly overvaluing features for grade 4 (even when the CNN correctly classified other ripeness stages). The situation remains unchanged when I visually examine the explanations generated by the methods across the entire dataset. This does not necessarily mean that the explanations generated by SHAP are not faithful to the model's inner workings. Rather, this work's intuitive interpretation of the highlighted regions is misaligned with how the model uses those features internally. As such, I can only conclude that, despite their plausibility, these visualizations are inadequate as significant explanations. However, this may not always hold. Assuming the visualization accurately reflects the model's internal workings, the positions of the key pixels would be much more meaningful for coarse-grained classification tasks. For instance, in a detection task on ImageNet, if the important pixels are located on an animal's muzzle, the intuitive interpretation is that the muzzle is the crucial part of the image. Nevertheless, even in this scenario, the interpretation of the heatmap is up to the user, not the model. In slightly more challenging cases, different users might arrive at different interpretations. This exemplifies why I believe saliency-based methods do not fully solve the explanation problem on their own but rather support the process of generating an explanation. In the specific dataset presented, however, the positions of these highlighted regions do not convey intelligible information. Conversely, the explanation for the DT is designed to be faithful to the model's inner workings. This strategy provides the user with a more informative, intuitively understandable, plausible, and faithful explanation of how the model works. An *ad hoc* user study confirms such results.

### 8.8.3 User Study

I designed a user study to investigate users' preferences for the model-generated explanations of model predictions. The users involved in the study are stakeholders in banana ripeness grading, comprising 20 people with diverse backgrounds and expertise in AI tools. I submitted an online questionnaire to each user. The complete questionnaire is reported in the supplementary material. The questionnaire introduces the task and asks the users to compare two types of explanations for the same input and prediction: (i) the mask generated by SHAP and (ii) the representation of the input color in the RGB gamut. Explanations (i) pertain to the ViT model (the best-performing one), while explanation (ii) is generated from the DT. The object of the comparison is how much the proposed explanation allows you to answer why the model made that prediction. When asked about the importance of explaining the model's behavior, all participants believed that an associated explanation is necessary, with most considering it essential. Regarding the preferred explanation method, 10 out of 20 respondents considered the DT-generated RGB gamut to be the most effective, while 8 chose the SHAP heatmap explanation. Three declared that no explanation was helpful to them [2]. This result is certainly interesting; although SHAP's visualizations do not provide an unambiguous explanation, their visual nature was sufficient to make half of the participants deem them trustworthy in explaining the prediction. Finally, 80% of respondents stated that the chosen explanation would improve their trust in the model, and 70% are willing to trade approximately 5% of the classifier's accuracy for a more transparent, human-explainable decision-making process. Considering that the accuracy loss between the DT and the most accurate model is only around 2.5% for this classification task, which is well above human performance, there appears to be little reason to prefer the latter to the more explainable one.

## 8.9 Limitations

Using simple classifiers on a few manually extracted features can be much more problematic for more complex tasks, as it can severely limit model performance. Indeed, I do not argue that more transparent models should *always* be used; many cognitive tasks would be nearly impossible without the progress achieved through DL. For this task, I selected a simple strategy to provide a clear, intuitive explanation to non-ML-expert users, based on the average color of the entire image. This can be refined iteratively to incorporate more complex features while accounting for explainability. I plan to explore strategies for serving explanations with richer feature representations, such as considering pixel color distributions and sampling to obtain more accurate color representations. In line with the *explainability by design* principle, I plan to investigate the use of regularization strategies to improve the explainability of complex DL models. This topic has been explored, with a focus on robustness, which has been linked to explainability [157]. It would be interesting to explore whether and how adding constraints on the features extracted by NNs could yield more interpretable explanations for end users.

[2] One participant selected both the RGB explanation and the "neither" option.

## 8.10 Conclusions

The banana ripeness case study offers a concrete illustration of the PAMC in practice. Its results reveal that model performance alone cannot justify the escalation of complexity, particularly for tasks that are perceptually or conceptually simple for human agents. When the structure of the problem is transparent, the most epistemically responsible strategy is not to retrofit explanations onto opaque models, but to design the model itself as an intelligible artifact. In this sense, the PAMC serves both as a diagnostic and a prescriptive principle, prompting researchers to question whether the additional layers of abstraction introduced by deep models yield epistemic gains commensurate with their cognitive and operational costs. This work demonstrates that the trade-off between predictive accuracy and explainability is not always antagonistic. A well-calibrated decision tree operating on color-based features achieves accuracy levels competitive with those of deep neural architectures, while remaining faithful to the problem's structure and intelligible to its stakeholders. The resulting explanation is not an afterthought derived from post-hoc methods but an emergent property of the model's design—one that aligns with users' intuitive reasoning about color and ripeness. This reinforces a central insight of the PAMC: model complexity should be appropriate to the epistemic and practical demands of the task, rather than maximized for its own sake.

At a broader level, this study also illustrates how explanation by design can reframe the relationship between performance and understanding. Instead of conceiving explainability as an external constraint that follows modeling, it becomes a design variable — one that modulates model selection, feature representation, and interpretive communication. The resulting pipeline exemplifies the kind of methodological reflexivity called for by the unified framework proposed in this thesis: faithfulness is here operationalized as a property of model structure, intelligibility as a function of stakeholder understanding, and alignment as a correspondence between representational semantics and human reasoning. In practical terms, the lesson is simple but far-reaching: *not every task requires a transformer*. When the epistemic structure of the problem is shallow, simplicity is not a limitation but a form of fidelity. Overkilling simple tasks with opaque systems introduces not only unnecessary technical overhead but also epistemic opacity and ethical cost. The PAMC invites us, therefore, to calibrate the sophistication of our models to the complexity of the world they seek to represent — an act of methodological humility that is also an act of scientific rigor.

# 9

# Conclusion

This thesis began with a deceptively simple yet profound question: *what is an explanation in the landscape of contemporary AI?* As DL and FMs have grown in scale and capability, their internal opacity has intensified a fundamental epistemic tension: how can we claim to understand systems that so clearly outperform us in reasoning tasks, yet resist intelligible explanation? The ambition of this work has been to address this tension by reconciling the theoretical and empirical dimensions of explainability, reframing it as a constitutive element of intelligent systems rather than a mere technical supplement.

The first part of the thesis traced the conceptual foundations of XAI, revealing three research directions that define its contemporary landscape: the *causal*, the *mechanistic*, and the *generative*. Each provides a partial account of what it means to explain. Causal explanation secures inferential grounding; mechanistic explanation opens the black box of model internals; and generative explanation situates understanding within communicative and narrative contexts. Yet, none alone is sufficient. Explanation, as argued throughout this work, is not an intrinsic property of a model but an epistemic relation between the system, its internal representations, and the human mind. Understanding emerges only when these relations are made coherent across cognitive, technical, and ethical dimensions.

From this synthesis emerged a unified theoretical framework structured around three constitutive dimensions of explainability: intelligibility, alignment, and faithfulness. These dimensions delineate the space within which explanation can meaningfully connect model behavior to human understanding. Intelligibility concerns the cognitive and communicative accessibility of explanations; alignment, their ethical and pragmatic resonance with human goals and values; and faithfulness, their fidelity to the model's internal causal and representational structure. Together, these dimensions form what this thesis has called the *triple frontier* of XAI pursuit — a framework that redefines explanation as the dynamic

mediation between human interpretive capacities and machine representational depth.

The framework was then examined empirically across three high-stakes domains — medicine, security, and industry — each of which foregrounds a distinct facet of this triple frontier. In medical imaging, the question of intelligibility took precedence: explanation served as a form of communication, translating algorithmic saliency into clinically meaningful evidence. In smart contract analysis and security, alignment came to the fore: explanations were understood as acts of knowledge construction that transformed latent model knowledge into actionable human insight. In the industrial context, faithfulness proved central: explanation became an act of design and governance, in which transparency and accountability were embedded into the very architecture of decision systems. Across these domains, the trajectory from communication to knowledge to design revealed that explainability is not a static outcome but a practice of epistemic mediation — one that evolves with context, purpose, and risk.

Among the three frontiers, faithfulness emerged as the most conceptually demanding and empirically elusive. It represents the epistemic integrity of the explanatory process: without faithfulness, intelligibility risks collapsing into plausible storytelling, and alignment risks degenerating into persuasion. To render faithfulness measurable rather than assumed, this thesis introduced a test suite for its empirical evaluation, centered on temporal attention mechanisms. This protocol demonstrated that faithfulness can, and must, be subjected to empirical scrutiny, establishing a foundation for a more rigorous, evidence-based science of explanation. Just as performance metrics anchor claims of accuracy, faithfulness metrics are needed to anchor claims of understanding.

From this realization emerged the final theoretical contribution: the *PAMC*. This principle asserts that model complexity should be viewed not as an unquestioned virtue, but as a parameter to be managed in relation to human comprehension and contextual needs. Complexity, in this sense, is neither an enemy of transparency nor an unqualified good; it is a variable that determines how explanation and understanding can coexist. The principle redefines the classic trade-off between accuracy and explainability as a question of proportionality — what level of complexity is *appropriate* for a given epistemic and ethical context. It calls for a mode of complexity governance in which the pursuit of performance is balanced by the pursuit of intelligibility, ensuring that systems deployed in high-stakes settings remain at a level commensurate with the transparency required by their use.

In the context of LLMs, this principle takes on renewed urgency. Emergent abilities and scale-dependent phenomena challenge the very notion of explainability, suggesting that complexity cannot be reduced without consequences. Yet, this does not imply resignation to opacity. Rather, it calls for new design strategies that render complexity legible through modular architectures, hybrid symbolic–statistical reasoning, and self-explanatory mechanisms. The goal is not to make complex systems simple, but to make their complexity *understandable*. In this light, explainability becomes an evolving property of the model ecosystem itself—a relation between compression, representation, and communication.

Taken together, the theoretical, empirical, and methodological contributions of this thesis converge on a broader insight: *explainability is not external to intelligence but intrinsic to it*. To explain is to establish the conditions under which reasoning can be shared, examined, and improved. Intelligence without explanation may be powerful, but it remains epistemically incomplete. The future of AI, therefore, depends not only on increasing computational capability but on cultivating systems capable of participating in the process of understanding itself.

# Limitations

Despite its integrative ambition, this thesis remains constrained by several important limitations. First, the empirical investigations, while diverse in domain, were necessarily limited in scope and scale. The test suite for faithfulness, for instance, focused on temporal attention mechanisms; extending this evaluation to multimodal and FMs would further validate its generality. Second, the proposed theoretical framework, though unifying in intent, does not fully capture the socio-technical dynamics that shape the interpretation of explanations in practice. Human understanding is always situated, and explanatory adequacy cannot be abstracted from cultural, institutional, and cognitive contexts. Finally, the PAMC, while conceptually grounded, remains an open challenge to operationalize quantitatively. Developing standardized measures of "appropriate" complexity—linking cognitive load, task criticality, and epistemic reliability—will require sustained interdisciplinary collaboration. These limitations do not weaken the argument but delineate the boundaries within which its claims should be interpreted and extended.

# Future Research Directions

The directions opened by this work point toward a redefinition of explainability as a field of co-evolving understanding between humans and machines. Future research should pursue three interrelated trajectories. First, integrating the unified framework into the study of FMs provides a path to empirically test how intelligibility, alignment, and faithfulness interact at scale, particularly in systems that generate natural-language explanations. Second, the development of self-explaining architectures—models that can introspect and justify their reasoning in real time—could transform explanation from an external intervention into an intrinsic cognitive function of the model itself. Third, human–AI co-adaptation should become a central focus: explanation should not only transmit information but also facilitate shared reasoning, enabling models to learn how different users conceptualize and validate their understanding. Advancing these directions would move the field beyond reactive explainability and toward proactive epistemic design.

# Concluding Remarks

The question that initiated this work—*what is an explanation in AI?*—thus transforms into a more profound one: *what does it mean for intelligence to be understood?* The answer, as developed throughout this thesis, lies in the dynamic balance between complexity and comprehension, performance and transparency, autonomy and accountability. To design explainable systems is therefore to design for understanding itself. It is to build bridges across the widening gulf between knowing and comprehending, between prediction and meaning, between intelligence and wisdom. In that endeavor, the measure of progress will not be the sophistication of our models, but the depth of our understanding of them, and ultimately, of ourselves.

# Declarations

## Use of Artificial Intelligence Tools

During the preparation of this thesis, the author employed Artificial Intelligence (AI) tools, including Large Language Models (LLMs), with the primary aim of improving language fluency, enhancing readability, and assisting with LaTeX formatting, code structuring, and reference management. All outputs and suggestions generated by these tools were critically evaluated, carefully edited, and fully validated by the author. The author affirms that the use of AI tools did not influence the scientific content, results, or interpretations presented herein. The responsibility for the originality, accuracy, and integrity of the final thesis rests solely with the author.

## Data and Code Availability

All software, experimental scripts, and evaluation frameworks developed as part of this thesis have been made openly available in public repositories to promote transparency, reproducibility, and further research. Details for accessing the data and code, including versioning and licensing information, are provided in the relevant sections of the thesis. Researchers are encouraged to use these resources in accordance with the specified licenses.

## Conflicts of Interest

The author declares that there are no known financial, personal, or professional relationships that could have influenced or appear to influence the research reported in this thesis. This declaration affirms the impartiality, objectivity, and integrity of the work presented.

# References


[1] Abdelhamid, M A and Sudnik, Y and Alshinayyin, H J and Shaaban, F. 'Nondestructive method for monitoring tomato ripening based on chlorophyll fluorescence induction'. In: *Journal of Agricultural Engineering Research* 52 (2021).

[2] Wael Abdelkader et al. 'Machine Learning Approaches to Retrieve High-Quality, Clinically Relevant Evidence From the Biomedical Literature: Systematic Review'. en. In: *JMIR Med Inform* 9.9 (Sept. 2021), e30401. ISSN: 2291-9694.

[3] Samira Abnar and Willem Zuidema. 'Quantifying Attention Flow in Transformers'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4190–4197.

[4] Radhakrishna Achanta et al. 'SLIC Superpixels Compared to State-of-the-Art Superpixel Methods'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34.11 (Nov. 2012), pp. 2274–2282. ISSN: 0162-8828.

[5] Amina Adadi and Mohammed Berrada. 'Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)'. In: *IEEE Access* 6 (Sept. 2018), pp. 52138–52160. ISSN: 2169-3536.

[6] Hajar Ait Addi, Redouane Ezzahir and Abdelhak Mahmoudi. 'Three-Level Binary Tree Structure for Sentiment Classification in Arabic Text'. In: *Proceedings of the 3rd International Conference on Networking, Information Systems & Security*. NISS2020. Marrakech, Morocco: Association for Computing Machinery, Mar. 2020, pp. 1–8. ISBN: 9781450376341.

[7] Julius Adebayo et al. 'Sanity Checks for Saliency Maps'. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Ed. by Samy Bengio et al. Vol. 31. NIPS'18. Montréal, Canada: Curran Associates Inc., Dec. 2018, pp. 9525–9536.

[8] Ashutosh Adhikari et al. 'DocBERT: BERT for Document Classification'. In: *arXiv (preprint)* abs/1904.08398 (May 2019). ISSN: 2331-8422. arXiv: 1904.08398 [cs.CL].

[9] Aghilinategh, Nahid and Dalvand, Mohammad Jafar and Anvar, Adieh. 'Detection of ripeness grades of berries using an electronic nose'. In: *Food Science & Nutrition* 8.9 (July 2020), pp. 4919–4928. ISSN: 2048-7177. (Visited on 21/02/2023).

[10] Neeraj Agrawal et al. 'Hierarchical Text Classification Using Contrastive Learning Informed Path Guided Hierarchy'. In: *ECAI 2023*. Ed. by Kobi Gal et al. Vol. 372. IOS Press, Sept. 2023, pp. 19–26. ISBN: 978-1643684369. (Visited on 10/01/2024).

[11] A. V. Aho, J. E. Hopcroft and J. D. Ullman. 'On Finding Lowest Common Ancestors in Trees'. In: *SIAM Journal on Computing* 5.1 (Mar. 1976), pp. 115–132. ISSN: 0097-5397.

[12] A. Albarelli et al. 'On the application of a common theoretical explainability framework in information retrieval'. In: *CEUR Workshop Proceedings*. Ed. by K. Roitero et al. Vol. 3802. CEUR-WS.org, 2024, pp. 43–52. URL: https://ceur-ws.org/Vol-3802/paper24.pdf.

[13] Nawal Aljedani, Reem Alotaibi and Mounira Taileb. 'HMATC: Hierarchical multi-label Arabic text classification model using machine learning'. In: *Egyptian Informatics Journal* 22.3 (Sept. 2021), pp. 225–237. ISSN: 1110-8665.

[14] Nawal Aljedani, Reem Alotaibi and Mounira Taileb. 'Multi-Label Arabic Text Classification: An Overview'. In: *International Journal of Advanced Computer Science and Applications* 11.10 (2020). ISSN: 2156-5570.

[15] Altaheri, Hamdi and Alsulaiman, Mansour and Muhammad, Ghulam. 'Date Fruit Classification for Robotic Harvesting in a Natural Environment Using Deep Learning'. In: *IEEE Access* 7 (Aug. 2019). Conference Name: IEEE Access, pp. 117115–117133. ISSN: 2169-3536.

[16] David Alvarez-Melis and Tommi S. Jaakkola. 'On the Robustness of Interpretability Methods'. In: *ICML Workshop on Human Interpretability in Machine Learning (WHI 2018)*. Vol. abs/1806.08049. Stockholm, Sweden: Cornell University, June 2018. arXiv: 1806.08049 [cs.LG].

[17] David Alvarez-Melis and Tommi S. Jaakkola. 'Towards robust interpretability with self-explaining neural networks'. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. Ed. by Samy Bengio et al. Vol. abs/1806.07538. NIPS'18. Montréal, Canada: Curran Associates Inc., June 2018, pp. 7786–7795.

[18] Rami Aly, Steffen Remus and Chris Biemann. ‘Hierarchical Multi-label Classification of Text with Capsule Networks’. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 2: Student Research Workshop*. Ed. by Fernando Alva-Manchego, Eunsol Choi and Daniel Khashabi. Association for Computational Linguistics, July 2019, pp. 323–330.

[19] Avinash Amballa et al. ‘Automated model selection for tabular data’. In: *arXiv preprint arXiv:2401.00961* abs/2401.00961 (Jan. 2024). ISSN: 2331-8422.

[20] Anzalone, Gerald C and Glover, Alexandra G and Pearce, Joshua M. ‘Open-source colorimeter’. In: *Sensors* 13.4 (May 2013), pp. 5338–5346. ISSN: 1424-8220.

[21] Laura Arbelaez Ossa et al. ‘Re-focusing explainability in medicine’. en. In: *Digit Health* 8 (Feb. 2022), p. 20552076221074488. ISSN: 2055-2076.

[22] Joaquín Arias et al. ‘Modeling and Reasoning in Event Calculus Using Goal-Directed Constraint Answer Set Programming’. In: *Springer*. Ed. by Maurizio Gabbrielli. Vol. 22. Springer, Oct. 2019, pp. 139–155.

[23] S. Ö Arık and T. Pfister. ‘TabNet: Attentive interpretable tabular learning’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 8. Association for the Advancement of Artificial Intelligence (AAAI), May 2021, pp. 6679–6687.

[24] *Assessment of what the consumer values in fresh fruit quality: Case study of Oman: New Zealand Journal of Crop and Horticultural Science: Vol 35, No 2*. (Visited on 21/02/2023).

[25] Jimmy Lei Ba, Jamie Ryan Kiros and Geoffrey E. Hinton. ‘Layer Normalization’. In: *arXiv* abs/1607.06450 (July 2016). ISSN: 2331-8422.

[26] Sebastian Bach et al. ‘On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation’. en. In: *PLoS One* 10.7 (July 2015), e0130140. ISSN: 1932-6203.

[27] Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. ‘Neural Machine Translation by Jointly Learning to Align and Translate’. In: *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Ed. by Yoshua Bengio and Yann LeCun. Vol. abs/1409.0473. Place: San Diego, CA, USA. ICLR 2015 oral session, 2015. arXiv: 1409.0473 [cs.CL].

[28] Baietto, Manuela and Wilson, Alphus D. ‘Electronic-nose applications for fruit identification, ripeness and quality grading’. In: *Sensors* 15.1 (Jan. 2015), pp. 899–931. ISSN: 1424-8220.

[29] Bakar, B and Ishak, A and Shamsudin, R and Hasan, W Wan. ‘Ripeness level classification for pineapple using rgb and hsi color map’. In: *Journal of Theoretical and Applied Information Technology* 57 (2013), pp. 587–593.

[30] Renu Balyan, Kathryn S. McCarthy and Danielle S. McNamara. ‘Applying Natural Language Processing and Hierarchical Machine Learning Approaches to Text Difficulty Classification’. In: *Int. J. Artif. Intell. Educ.* 30.3 (June 2020), pp. 337–370. ISSN: 1560-4292.

[31] Rajdeep Banerjee and Somesh Kr Bhattacharya. ‘Data: Periodicity and Ways to Unlock Its Full Potential’. In: *Rhythmic Advantages in Big Data and Machine Learning* (2022), pp. 1–22. ISSN: 2524-5546.

[32] Siddhartha Banerjee et al. ‘Hierarchical Transfer Learning for Multi-label Text Classification’. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Ed. by Anna Korhonen, David R. Traum and Lluís Màrquez. Association for Computational Linguistics, July 2019, pp. 6295–6300.

[33] Jinhyun Bang, Jonghun Park and Jonghyuk Park. ‘GACaps-HTC: graph attention capsule network for hierarchical text classification’. In: *Applied Intelligence* 53.17 (Sept. 2023), pp. 20577–20594. ISSN: 1573-7497.

[34] Bargoti, Suchet and Underwood, James. *Deep Fruit Detection dataset*. 2016.

[35] Bargoti, Suchet and Underwood, James. ‘Deep Fruit Detection in Orchards’. In: *2017 IEEE International Conference on Robotics and Automation (ICRA)*. Singapore, Singapore: IEEE Press, May 2017, pp. 3626–3633.

[36] Solon Barocas, Moritz Hardt and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. en. MIT Press, Dec. 2023.

[37] Jonathan T. Barron and Yun-Ta Tsai. ‘Fast Fourier Color Constancy’. In: *Cvpr*. IEEE Computer Society, July 2017, pp. 6950–6958.

[38] Bauriegel, E and Giebel, A and Geyer, M and Schmidt, U and Herppich, W B. ‘Early detection of Fusarium infection in wheat using hyper-spectral imaging’. In: *Computers and Electronics in Agriculture* 75.2 (Feb. 2011). Publisher: Elsevier BV, pp. 304–312. ISSN: 0168-1699.

[39] Vaishak Belle and Ioannis Papantonis. ‘Principles and Practice of Explainable Machine Learning’. en. In: *Front Big Data* 4 (July 2021), p. 688969. ISSN: 2624-909X.

[40] El-Bendary, Nashwa and El Hariri, Esraa and Hassanien, Aboul Ella and Badr, Amr. ‘Using machine learning techniques for evaluating tomato ripeness’. In: *Expert Systems with Applications* 42.4 (Mar. 2015), pp. 1892–1905. ISSN: 0957-4174. (Visited on 21/02/2023).

[41] Yoav Benjamini and Yosef Hochberg. ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 57.1 (Jan. 1995), pp. 289–300. ISSN: 00359246.

[42] Rohan Bhambhoria, Lei Chen and Xiaodan Zhu. ‘A Simple and Effective Framework for Strict Zero-Shot Hierarchical Classification’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*. Ed. by Anna Rogers, Jordan L. Boyd-Graber and Naoaki Okazaki. Vol. abs/2305.15282. Association for Computational Linguistics, May 2023, pp. 1782–1792.

[43] S. Bianco and C. Cusano. ‘Quasi-Unsupervised Color Constancy’. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 12204–12213.

[44] Francesco Bianconi, Jakob N. Kather and Constantino Carlos Reyes-Aldasoro. ‘Experimental Assessment of Color Deconvolution and Color Normalization for Automated Classification of Histology Images Stained with Hematoxylin and Eosin’. In: *Cancers* 12.11 (Nov. 2020), p. 3337. ISSN: 2072-6694.

[45] Alexander Binder et al. ‘Layer-Wise Relevance Propagation for Neural Networks with Local Renormalization Layers’. In: *Artificial Neural Networks and Machine Learning – ICANN 2016*. Ed. by Alessandro E.P. Villa, Paolo Masulli and Antonio Javier Pons Rivero. Vol. abs/1604.00825. Cham: Springer International Publishing, Apr. 2016, pp. 63–71. ISBN: 978-3-319-44781-0.

[46] Bodria, L and Fiala, M and Guidetti, R and Oberti, R. ‘Optical techniques to estimate the ripeness of red-pigmented fruits’. In: *Transactions of the ASAE* 47.3 (2004), pp. 815–820. ISSN: 0001-2351.

[47] Biagio Boi, Christian Esposito and Sokjoon Lee. ‘Smart Contract Vulnerability Detection: The Role of Large Language Model (LLM)’. In: *SIGAPP Appl. Comput. Rev.* 24.2 (June 2024), pp. 19–29. ISSN: 1559-6915.

[48] Mathias Bollaert, Olivier Augereau and Gilles Coppin. ‘Measuring and Calibrating Trust in Artificial Intelligence’. In: *Lecture Notes in Computer Science*. Springer Nature Switzerland, Mar. 2024, pp. 232–237. ISBN: 978-3031616976.

[49] Lorenzo Bongiovanni et al. ‘Zero-Shot Taxonomy Mapping for Document Classification’. In: *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing*. Ed. by Jiman Hong et al. SAC ’23. Tallinn, Estonia: Association for Computing Machinery, Mar. 2023, pp. 911–918. ISBN: 9781450395175.

[50] Bonora, E and Vidoni, S and Noferini, M and Costa, G and Lopresti, J and Stefanelli, D. ‘The combined use of the index of absorbance difference and the reconstruction model Plantoon® to characterize peach and nectarine training systems’. In: *Acta Horticulturae* 1084 (2015), pp. 361–366.

[51] Antoine Bordes et al. ‘Translating Embeddings for Modeling Multi-relational Data’. In: *Advances in Neural Information Processing Systems*. Ed. by C.J. Burges et al. Vol. 26. Curran Associates, Inc., Dec. 2013.

[52] Antal van den Bosch. ‘Hidden Markov Models’. In: *Encyclopedia of Machine Learning and Data Mining*. Boston, MA: Springer US, 2017. Chap. Hidden Markov Models, pp. 609–611. ISBN: 978-1-4899-7687-1.

[53] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. ‘A Training Algorithm for Optimal Margin Classifiers’. In: *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*. Association for Computing Machinery, July 1992, pp. 144–152. ISBN: 089791497X.

[54] Leo Breiman. *Classification and Regression Trees*. Vol. 40. 3. New York: Routledge, Sept. 1984, p. 874. ISBN: 9781315139470.

[55] Leo Breiman. ‘Random forests’. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 0885-6125.

[56] Tom Brown et al. ‘Language Models are Few-Shot Learners’. In: *Advances in Neural Information Processing Systems*. Ed. by Hugo Larochelle et al. Vol. 33. Curran Associates, Inc., May 2020, pp. 1877–1901.

[57] Sébastien Bubeck et al. ‘Sparks of Artificial General Intelligence: Early experiments with GPT-4’. Mar. 2023.

[58] Enrico Bugiardini et al. ‘The diagnostic value of MRI pattern recognition in distal myopathies’. In: *Frontiers in neurology* 9 (June 2018), p. 456. ISSN: 1664-2295.

[59] Bulanon, D and Burks, T and Alchanatis, V. *Visible and Thermal Images for Fruit Detection*. Pages: 944–954. 2011.

[60] Danielle Caled et al. ‘A Hierarchical Label Network for Multi-label EuroVoc Classification of Legislative Contents’. In: *Digital Libraries for Open Knowledge*. Ed. by Antoine Doucet et al. Vol. 11799. Cham: Springer International Publishing, Sept. 2019, pp. 238–252. ISBN: 978-3-030-30760-8.

[61] Camps, C and Christen, D. ‘Non-destructive assessment of apricot fruit quality by portable visible-near infrared spectroscopy’. In: *LWT - Food Science and Technology* 42.6 (July 2009), pp. 1125–1131. ISSN: 0023-6438.

[62] Gianluca Carloni, Andrea Berti and Sara Colantonio. *The role of causality in explainable artificial intelligence*. Sept. 2023.

[63] Luciano Caroprese, Eugenio Vocaturo and Ester Zumpano. ‘Argumentation approaches for explanaible AI in medical informatics’. In: *Intelligent Systems with Applications* 16 (Nov. 2022), p. 200109. ISSN: 2667-3053. (Visited on 28/05/2024).

[64] Rich Caruana et al. ‘Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission’. In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by Longbing Cao et al. KDD ’15. Sydney, NSW, Australia: Association for Computing Machinery, Aug. 2015, pp. 1721–1730. ISBN: 9781450336642.

[65] Carvalho, Diogo V and Pereira, Eduardo M and Cardoso, Jaime S. ‘Machine learning interpretability: A survey on methods and metrics’. In: *Electronics (Basel)* 8.8 (July 2019). Publisher: MDPI AG, p. 832. ISSN: 2079-9292.

[66] Castro, Wilson and Oblitas, Jimy and De-La-Torre, Miguel and Cotrina, Carlos and Bazan, Karen and Avila-George, Himer. ‘Classification of Cape Gooseberry Fruit According to its Level of Ripeness Using Machine Learning Techniques and Different Color Spaces’. In: *IEEE Access* 7 (Mar. 2019). Publisher: Institute of Electrical and Electronics Engineers (IEEE), pp. 27389–27400. ISSN: 2169-3536.

[67] Michelangelo Ceci and Donato Malerba. ‘Classifying web documents in a hierarchy of categories: a comprehensive study’. In: *Journal of Intelligent Information Systems* 28.1 (Feb. 2007), pp. 37–78. ISSN: 1573-7675.

[68] Centner, V and Massart, D L and de Noord, O E and de Jong, S and Vandeginste, B M and Sterna, C. ‘Elimination of uninformative variables for multivariate calibration’. In: *Analytical Chemistry* 68.21 (Nov. 1996). Publisher: American Chemical Society, pp. 3851–3858. ISSN: 0003-2700. (Visited on 21/02/2023).

[69] Jurgita Černevičienė and Audrius Kabašinskas. ‘Explainable artificial intelligence (XAI) in finance: a systematic literature review’. In: *Artificial Intelligence Review* 57.8 (July 2024), p. 216. ISSN: 0269-2821.

[70] Ricardo Cerri, Rodrigo C. Barros and André C. P. L. F. de Carvalho. ‘Hierarchical multi-label classification for protein function prediction: A local approach based on neural networks’. In: *2011 11th International Conference on Intelligent Systems Design and Applications*. Ed. by Sebastián Ventura et al. Vol. 1. IEEE, Nov. 2011, pp. 337–343.

[71] Ricardo Cerri et al. ‘Inducing Hierarchical Multi-label Classification rules with Genetic Algorithms’. In: *Applied Soft Computing* 77 (May 2019), pp. 584–604. ISSN: 1568-4946.

[72] Ilias Chalkidis et al. ‘An Empirical Study on Large-Scale Multi-Label Text Classification Including Few and Zero-Shot Labels’. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Bonnie Webber et al. Online: Association for Computational Linguistics, Nov. 2020, pp. 7503–7515.

[73] Chun Sik Chan, Huanqi Kong and Liang Guanqing. ‘A Comparative Study of Faithfulness Metrics for Model Interpretability Methods’. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 5029–5038.

[74] Lawrence Chan et al. ‘Causal scrubbing, a method for rigorously testing interpretability hypotheses’. In: *AI Alignment Forum* (2022). https://www.alignmentforum.org/posts/JvZhhzycHu2Yd57RN/causal-scrubbing-a-method-for-rigorously-testing.

[75] Soumya Chatterjee et al. ‘Joint Learning of Hyperbolic Label Embeddings for Hierarchical Multi-label Classification’. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Ed. by Paola Merlo, Jörg Tiedemann and Reut Tsarfaty. Vol. abs/2101.04997. Online: Association for Computational Linguistics, May 2021, pp. 2829–2841.

[76] Aditya Chattopadhay et al. ‘Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks’. In: *2018 IEEE winter conference on applications of computer vision (WACV)*. IEEE. IEEE Computer Society, Mar. 2018, pp. 839–847.

[77] Chawla, N V and Bowyer, K W and Hall, L O and Kegelmeyer, W P. ‘SMOTE: Synthetic Minority Over-sampling Technique’. In: *jair* 16 (June 2002), pp. 321–357. ISSN: 1076-9757.

[78] Larissa Chazette and Kurt Schneider. ‘Explainability as a non-functional requirement: challenges and recommendations’. en. In: *Requirements Engineering* 25.4 (Dec. 2020), pp. 493–514. ISSN: 0947-3602.

[79] Anni Chen and Bhuwan Dhingra. ‘Hierarchical Multi-Instance Multi-Label Learning for Detecting Propaganda Techniques’. In: *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*. Ed. by Burcu Can et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 155–163.

[80] Boli Chen et al. ‘Hyperbolic Interaction Model for Hierarchical Multi-Label Classification’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 7496–7503. ISSN: 2159-5399.

[81] Chacha Chen et al. ‘Machine Explanations and Human Understanding’. In: *arXiv (Cornell University)* (Feb. 2022). arXiv: 2202.04092 [cs.AI].

[82] Chaofan Chen et al. ‘This Looks like That: Deep Learning for Interpretable Image Recognition’. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Ed. by Hanna M. Wallach et al. Vol. 32. Red Hook, NY, USA: Curran Associates Inc., 2019, pp. 8928–8939.

[83] Chih Yao Chen et al. ‘Label-Aware Hyperbolic Embeddings for Fine-grained Emotion Classification’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Vol. abs/2306.14822. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 10947–10958.

[84] Haibin Chen et al. ‘Hierarchy-aware Label Semantics Matching Network for Hierarchical Text Classification’. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong et al. Association for Computational Linguistics, 2021, pp. 4370–4379.

[85] Jie Chen et al. ‘CLEP: A Novel Contrastive Learning Method for Evolutionary Reentrancy Vulnerability Detection’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Ed. by Toby Walsh, Julie Shah and Zico Kolter. Vol. 39. 1. Association for the Advancement of Artificial Intelligence (AAAI), May 2025, pp. 67–74.

[86] Jie Chen et al. ‘Research on patent classification based on hierarchical label semantics’. In: *2022 3rd International Conference on Education, Knowledge and Information Management (ICEKIM)*. IEEE, Jan. 2022, pp. 1025–1032.

[87] Kuan-Chun Chen, Cheng-Te Li and Kuo-Jung Lee. ‘DDNAS: Discretized Differentiable Neural Architecture Search for Text Classification’. In: *ACM Trans. Intell. Syst. Technol.* 14.5 (Oct. 2023), pp. 1–22. ISSN: 2157-6904.

[88] Lei Chen, Houwei Chou and Xiaodan Zhu. ‘Developing Prefix-Tuning Models for Hierarchical Text Classification’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: Industry Track*. Ed. by Yunyao Li and Angeliki Lazaridou. Abu Dhabi, UAE: Association for Computational Linguistics, Dec. 2022, pp. 390–397.

[89] Y. Chen et al. ‘Learning deep representations for the multi-branch neural networks’. In: *IEEE Transactions on Neural Networks and Learning Systems* 30.8 (2019), pp. 2396–2406.

[90] Dongliang Cheng, D. Prasad and M. Brown. ‘Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution.’ In: *Journal of the Optical Society of America. A, Optics, image science, and vision* 31 5.5 (May 2014), pp. 1049–58. ISSN: 1084-7529.

[91] Quan Cheng and Yingru Lin. 'Multilevel Classification of Users' Needs in Chinese Online Medical and Health Communities: Model Development and Evaluation Based on Graph Convolutional Network'. In: *JMIR Form Res* 7 (May 2023), e42297. ISSN: 2561-326X.

[92] Valeriia Cherepanova et al. 'A performance-driven benchmark for feature selection in tabular deep learning'. In: *Advances in Neural Information Processing Systems* 36 (Nov. 2023). Ed. by Alice Oh et al., pp. 41956–41979.

[93] Kyunghyun Cho et al. 'Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation'. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. Association for Computational Linguistics, Oct. 2014, pp. 1724–1734.

[94] Kyunghyun Cho et al. 'On the Properties of Neural Machine Translation: Encoder–Decoder Approaches'. In: *Proceedings of SSST-8, Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation*. Ed. by Dekai Wu et al. Doha, Qatar: Association for Computational Linguistics, Oct. 2014, pp. 103–111.

[95] Edward Choi et al. 'RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism'. en. In: *Advances in Neural Information Processing Systems* 29 (Aug. 2016). Ed. by Daniel D. Lee et al., pp. 3504–3512. (Visited on 31/05/2021).

[96] Francois Chollet. 'Xception: Deep learning with depthwise separable convolutions'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, July 2017, pp. 1800–1807.

[97] Choo, Wee Sim. 'Fruit Pigment Changes During Ripening'. In: *Encyclopedia of Food Chemistry*. Ed. by Laurence Melton, Fereidoon Shahidi and Peter Varelis. Oxford: Academic Press, Jan. 2019, pp. 117–123. ISBN: 978-0-12-814045-1. (Visited on 21/02/2023).

[98] Molnar Christoph. *Interpretable machine learning: A guide for making black box models explainable*. Leanpub, 2020.

[99] Michael Chromik and Andreas Butz. 'Human-XAI Interaction: A Review and Design Principles for Explanation User Interfaces'. In: *Human-Computer Interaction – INTERACT 2021*. Ed. by Carmelo Ardito et al. Vol. 12933. Springer International Publishing, 2021, pp. 619–640. ISBN: 978-3030856151.

[100] Bilal Chughtai, Lawrence Chan and Neel Nanda. 'A toy model of universality: reverse engineering how networks learn group operations'. In: *Proceedings of the 40th International Conference on Machine Learning*. Ed. by Andreas Krause et al. Vol. abs/2302.03025. ICML'23. Honolulu, Hawaii, USA: JMLR.org, Feb. 2023.

[101] Andrea Ciapetti et al. 'NETHIC: A System for Automatic Text Classification using Neural Networks and Hierarchical Taxonomies'. In: *Proceedings of the 21st International Conference on Enterprise Information Systems, ICEIS 2019, Heraklion, Crete, Greece, May 3-5, 2019, Volume 1*. Ed. by Joaquim Filipe et al. SciTePress, 2019, pp. 296–306.

[102] Florian Ciurea and Brian Funt. 'A Large Image Database for Color Constancy Research'. In: *Imaging Science and Technology Eleventh Color Imaging Conference*. Vol. 11. 1. Society for Imaging Science & Technology, Jan. 2003, pp. 160–164.

[103] Miruna-Adriana Clinciu and Helen Hastie. 'A Survey of Explainable AI Terminology'. In: *Proceedings of the 1st Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence (NL4XAI 2019)*. Ed. by Jose M Alonso and Alejandro Catala. Association for Computational Linguistics, 2019, pp. 8–13.

[104] Carlo Combi et al. 'A manifesto on explainability for artificial intelligence in medicine'. en. In: *Artif. Intell. Med.* 133 (Nov. 2022), p. 102423. ISSN: 0933-3657.

[105] ConsenSys. *Mythril: Symbolic-Execution-Based Security Analysis Tool for EVM Bytecode*. `https://github.com/ConsenSysDiligence/mythril`. [Accessed 30-05-2025].

[106] GitHub Contributors. *GitHub - xf97/HuangGai at v1.0.0*. `https://github.com/xf97/HuangGai/tree/v1.0.0`. [Accessed 29-05-2025].

[107] Alana de Santana Correia and Esther Luna Colombini. 'Attention, please! A survey of Neural Attention Models in Deep Learning'. In: *arXiv:2103.16775 [cs]* abs/2103.16775 (Mar. 2021). arXiv: 2103.16775. (Visited on 31/05/2021).

[108] Corinna Cortes and Vladimir Naumovich Vapnik. 'Support-Vector Networks'. In: *Machine Learning.* 20.3 (Sept. 1995), pp. 273–297. ISSN: 0885-6125.

[109] Costa, G and Noferini, M and Fiori, G and Ziosi, V and Berthod, N and Rossier, J. 'Establishment of the optimal harvest time in apricot ('orangered' and 'bergarouge') by means of a new index based on vis spectroscopy'. In: *Acta Horticulturae* 862 (May 2010), pp. 533–539. ISSN: 0567-7572.

[110] Ian C. Covert, Scott Lundberg and Su-In Lee. 'Explaining by removing: a unified framework for model explanation'. In: *J. Mach. Learn. Res.* 22.1 (Jan. 2021). ISSN: 1532-4435.

[111] Joao Crisostomo, Fernando Bacao and Victor Lobo. 'Machine learning methods for detecting smart contracts vulnerabilities within Ethereum blockchain - A review'. In: *Expert Systems with Applications* 268 (May 2025), p. 126353. ISSN: 0957-4174.

[112] Marina Danilevsky et al. 'A Survey of the State of Explainable AI for Natural Language Processing'. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*. Ed. by Kam-Fai Wong, Kevin Knight and Hua Wu. Vol. abs/2010.00711. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 447–459.

[113] Das, A J and Wahi, A and Kothari, I and Raskar, R. 'Ultra-portable, wireless smartphone spectrometer for rapid, non-destructive testing of fruit ripeness'. In: *Scientific Reports* 6.1 (Sept. 2016), p. 32504. ISSN: 2045-2322.

[114] *DBpedia: Home*. Dec. 2023. (Visited on 22/01/2024).

[115] M. Defferrard, X. Bresson and P. Vandergheynst. 'Convolutional neural networks on graphs with fast localized spectral filtering'. In: *Advances in Neural Information Processing Systems*. Ed. by Daniel D. Lee et al. Vol. 29. Cornell University, June 2016, pp. 3837–3845.

[116] Sofi Defiyanti, Edi Winarko and Sigit Priyanta. 'A Survey of Hierarchical Classification Algorithms with Big-Bang Approach'. In: *2019 5th International Conference on Science and Technology (ICST)*. Vol. 1. IEEE, July 2019, pp. 1–6.

[117] J. Deng et al. 'ImageNet: A Large-Scale Hierarchical Image Database'. In: *Cvpr09*. IEEE, June 2009, pp. 248–255.

[118] Li Deng. 'The mnist database of handwritten digit images for machine learning research'. In: *IEEE Signal Processing Magazine* 29.6 (Nov. 2012), pp. 141–142. ISSN: 1053-5888.

[119] Zhongfen Deng et al. 'HTCInfoMax: A Global Model for Hierarchical Text Classification via Information Maximization'. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Ed. by Kristina Toutanova et al. Association for Computational Linguistics, May 2021, pp. 3259–3265.

[120] Jacob Devlin et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186.

[121] Jay DeYoung et al. 'ERASER: A Benchmark to Evaluate Rationalized NLP Models'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4443–4458.

[122] Dhanoa, M S and Lister, S J and Sanderson, R and Barnes, R J. 'The link between multiplicative scatter correction (MSC) and standard normal variate (SNV) transformations of NIR spectra'. In: *Journal of Near Infrared Spectroscopy* 2.1 (Jan. 1994). Publisher: SAGE Publications, pp. 43–47. ISSN: 0967-0335.

[123] Michele Di Angelo and Gernot Salzer. 'A Survey of Tools for Analyzing Ethereum Smart Contracts'. In: *2020 IEEE International Conference on Decentralized Applications and Infrastructures (DAPPCON)*. IEEE, 2020, pp. 115–120.

[124] Monika Di Angelo and Gernot Salzer. 'Consolidation of Ground Truth Sets for Weakness Detection in Smart Contracts'. In: *Financial Cryptography and Data Security. FC 2023 International Workshops - Voting, CoDecFin, DeFi, WTSC, Bol, Brač, Croatia, May 5, 2023, Revised Selected Papers*. Ed. by Aleksander Essex et al. Vol. 13953. Springer Science+Business Media, Dec. 2023, pp. 439–455. ISBN: 978-3-031-48805-4.

[125] Monika di Angelo et al. 'SmartBugs 2.0: An Execution Framework for Weakness Detection in Ethereum Smart Contracts'. In: *38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*. IEEE Computer Society, Sept. 2023, pp. 2102–2105.

[126] Josep Domingo-Ferrer et al. *Tabular Data.* 2018.

[127] Guozhong Dong et al. 'OWGC-HMC: An Online Web Genre Classification Model Based on Hierarchical Multilabel Classification'. In: *Security and Communication Networks* 2022 (Mar. 2022), p. 7549880. ISSN: 1939-0114.

[128] Hang Dong et al. 'Automated Social Text Annotation With Joint Multilabel Attention Networks'. In: *IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS* 32.5 (May 2021), pp. 2224–2238. ISSN: 2162-237X.

[129] Finale Doshi-Velez and Been Kim. 'Towards a rigorous science of interpretable machine learning'. In: *arXiv preprint arXiv:1702.08608* (Feb. 2017).

[130] Alexey Dosovitskiy et al. 'An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale'. In: *International Conference on Learning Representations*. OpenReview.net, May 2021.

[131] Rachel Lea Draelos and Lawrence Carin. 'Use HiResCAM instead of Grad-CAM for faithful explanations of convolutional neural networks'. In: *arXiv preprint arXiv:2011.08891* (Nov. 2020).

[132] Maximilian Dreyer et al. 'Mechanistic understanding and validation of large AI models with SemanticLens'. en. In: *Nature Machine Intelligence* 7 (Aug. 2025). Publisher: Nature Publishing Group, pp. 1–14. ISSN: 2522-5839. (Visited on 18/08/2025).

[133] Xueying Du et al. 'Evaluating Large Language Models in Class-Level Code Generation'. In: *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*. ICSE '24. Lisbon, Portugal: Association for Computing Machinery, May 2024, pp. 1–13. ISBN: 9798400702174.

[134] Thomas Durieux et al. 'Empirical Review of Automated Analysis Tools on 47,587 Ethereum Smart Contracts'. In: *Proceedings of the ACM/IEEE 42nd International conference on software engineering*. Ed. by Gregg Rothermel and Doo-Hwan Bae. ACM, June 2020, pp. 530–541.

[135] Phan The Duy et al. 'Vulnsense: Efficient vulnerability detection in ethereum smart contracts by multimodal learning with graph neural network and language model'. In: *International Journal of Information Security* 24.1 (Feb. 2025), p. 48. ISSN: 1615-5262.

[136] EC. *LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS*. en. 2021. (Visited on 28/09/2023).

[137] D. Edgar, R. Biloslavo and M. Rizzo. 'The role of artificial intelligence in change management: Opportunities and challenges'. In: *Journal of Organizational Change Management* (2025). Manuscript accepted for publication.

[138] Nelson Elhage et al. 'A Mathematical Framework for Transformer Circuits'. In: *Transformer Circuits Thread* (2021). https://transformer-circuits.pub/2021/framework/index.html.

[139] Jeffrey L. Elman. 'Finding Structure in Time'. In: *Cognitive Science* 14.2 (Mar. 1990), pp. 179–211. ISSN: 0364-0213.

[140] ElMasry, G M and Nakauchi, S. ‘Image analysis operations applied to hyperspectral images for non-invasive sensing of food quality – a comprehensive review’. In: *Biosystems Engineering* 142 (Feb. 2016), pp. 53–82. ISSN: 1537-5110.

[141] ElMasry, Gamal and Wang, Ning and ElSayed, Adel and Ngadi, Michael. ‘Hyperspectral imaging for nondestructive determination of some quality attributes for strawberry’. In: *Journal of Food Engineering* 81.1 (July 2007). Publisher: Elsevier BV, pp. 98–107. ISSN: 0260-8774.

[142] Glyn Elwyn et al. ‘Shared decision making: a model for clinical practice’. en. In: *J. Gen. Intern. Med.* 27.10 (Oct. 2012), pp. 1361–1367. ISSN: 0884-8734.

[143] Adrian Erasmus, Tyler D P Brunet and Eyal Fisher. ‘What is Interpretability?’ en. In: *Philos. Technol.* 34.4 (Dec. 2021), pp. 833–862. ISSN: 2210-5433.

[144] Vincent Fabry et al. ‘A deep learning tool without muscle-by-muscle grading to differentiate myositis from facio-scapulohumeral dystrophy using MRI’. In: *Diagnostic and Interventional Imaging* 103.7-8 (July 2022), pp. 353–359. ISSN: 2211-5684.

[145] Matús Falis et al. ‘CoPHE: A Count-Preserving Hierarchical Evaluation Metric in Large-Scale Multi-Label Text Classification’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, Sept. 2021, pp. 907–912.

[146] Qingwu Fan and Changsheng Qiu. ‘Hierarchical Multi-label Text Classification Method Based On Multi-level Decoupling’. In: *2023 3rd International Conference on Neural Networks, Information and Communication Engineering (NNICE)*. IEEE, Feb. 2023, pp. 453–457.

[147] FDA. *Using Artificial Intelligence & Machine Learning in the Development of Drugs & Biological Products*. Online. 2021.

[148] Fda. ‘FDA Regulation of AI and Machine Learning in Medical Devices’. In: (2021).

[149] William Fedus, Barret Zoph and Noam Shazeer. ‘Switch Transformers: Scaling to Trillion Parameter Models with Simple and Efficient Sparsity’. In: *Journal of Machine Learning Research* 23.120 (Feb. 2022), pp. 1–39.

[150] Josselin Feist, Gustavo Grieco and Alexandru Munteanu. ‘Slither: A static analysis framework for smart contracts’. In: *Proceedings of the 2nd International Workshop on Emerging Trends in Software Engineering for Blockchain*. IEEE / ACM, May 2019, pp. 8–15.

[151] Paolo Florent Felisaz et al. ‘Texture analysis and machine learning to predict water T2 and fat fraction from non-quantitative MRI of thigh muscles in Facioscapulohumeral muscular dystrophy’. In: *European journal of radiology* 134 (Jan. 2021), p. 109460. ISSN: 0720-048X.

[152] Sohrab Ferdowsi et al. ‘Classification of hierarchical text using geometric deep learning: the case of clinical trials corpus’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Vol. abs/2110.15710. Association for Computational Linguistics, Oct. 2021, pp. 608–618.

[153] Ferrer, A and Remon, S and Negueruela, A I and Oria, R. ‘Changes during the ripening of the very late season spanish peach cultivar calanda: Feasibility of using cielab coordinates as maturity indices’. In: *Scientia Horticulturae* 105 (2005), pp. 435–446.

[154] Andy Field. *Discovering Statistics Using IBM SPSS Statistics*. 4th. Sage Publications Ltd., 2013. ISBN: 1446249182.

[155] Elena Fitkov-Norris, Samireh Vahid and Chris Hand. ‘Evaluating the impact of categorical data encoding and scaling on neural network classification performance: The case of repeat consumption of identical cultural goods’. In: *Engineering Applications of Neural Networks: 13th International Conference, EANN 2012, London, UK, September 20-23, 2012. Proceedings 13*. Ed. by Chrisina Jayne, Shigang Yue and Lazaros S. Iliadis. Vol. 311. Springer. Springer Science+Business Media, Sept. 2012, pp. 343–352. ISBN: 978-3642329081.

[156] Luciano Floridi. ‘A Conjecture on a Fundamental Trade-off between Certainty and Scope in Symbolic and Generative AI’. In: *SSRN Electronic Journal* abs/2506.10130 (June 2025). Available at SSRN: `https://ssrn.com/abstract=5289884` or `http://dx.doi.org/10.2139/ssrn.5289884`. ISSN: 1556-5068.

[157] David H. Foster. ‘Color constancy’. In: *Vision Research* 51.7 (May 2011). Vision Research 50th Anniversary Issue: Part 1, pp. 674–700. ISSN: 0042-6989.

[158] G. Frasson et al. ‘Assessing the value of explainable artificial intelligence for magnetic resonance imaging’. In: *Explainable artificial intelligence. xAI 2025*. Ed. by R. Guidotti, U. Schmid and L. Longo. Vol. 2576. Communications in Computer and Information Science. Cham: Springer, 2026, pp. 320–334. ISBN: 978-3032083166. DOI: `10.1007/978-3-032-08317-3_20`.

[159] Alex Freitas and Andre de Carvalho. ‘A Tutorial on Hierarchical Classification with Applications in Bioinformatics’. In: *Research and Trends in Data Mining Technologies and Applications* (Jan. 2007), pp. 175–208.

[160] J. H. Friedman. ‘Greedy function approximation: A gradient boosting machine’. In: *Annals of Statistics* 29.5 (Oct. 2001), pp. 1189–1232. ISSN: 0090-5364.

[161] Jun Fu et al. ‘Dual Attention Network for Scene Segmentation’. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 3146–3154. (Visited on 31/05/2021).

[162] Caro Fuchs et al. ‘pyFUME: a Python Package for Fuzzy Model Estimation’. In: *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. Glasgow, United Kingdom: IEEE, July 2020, pp. 1–8.

[163] Caro Fuchs et al. ‘The Impact of Variable Selection and Transformation on the Interpretability and Accuracy of Fuzzy Models’. In: *2022 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Ottawa, ON, Canada: IEEE, Aug. 2022, pp. 1–8.

[164] Joseph Futoma, Sanjay Hariharan and Katherine Heller. ‘Learning to Detect Sepsis with a Multitask Gaussian Process RNN Classifier’. In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1174–1182.

[165] Sandeep Gantla, Victor Ogunrinde and Kushal Gurram. ‘Exploring Mechanistic Interpretability in Large Language Models: Challenges, Approaches, and Insights’. In: (Mar. 2025), pp. 28–29.

[166] Gao, Zongmei and Shao, Yuanyuan and Xuan, Guantao and Wang, Yongxian and Liu, Yi and Han, Xiang. ‘Real-time hyperspectral imaging for the in-field estimation of strawberry ripeness with deep learning’. In: *Artificial Intelligence in Agriculture* 4 (2020). Publisher: Elsevier BV, pp. 31–38. ISSN: 2589-7217.

[167] Francesco Gargiulo et al. ‘Deep neural network for hierarchical extreme multi-label text classification’. In: *Appl. Soft Comput.* 79 (June 2019), pp. 125–138. ISSN: 1568-4946.

[168] Garillos-Manliguez, Cinmayii A and Chiang, John Y. ‘Multimodal Deep Learning and Visible-Light and Hyperspectral Imaging for Fruit Maturity Estimation’. In: *Sensors* 21.4 (Feb. 2021), p. 1288. ISSN: 1424-8220.

[169] Damien Garreau and Dina Mardaoui. ‘What does LIME really see in images?’ In: *ICML 2021 - 38th International Conference on Machine Learning*. Ed. by Marina Meila and Tong Zhang. Vol. 139. Virtual Conference, United States: Cornell University, July 2021, pp. 3620–3629.

[170] Andrea Gasparetto, Giorgia Minello and Andrea Torsello. ‘Non-parametric Spectral Model for Shape Retrieval’. In: *2015 International Conference on 3D Vision*. Ed. by Michael S. Brown, Jana Kosecká and Christian Theobalt. Vol. 32. IEEE, Oct. 2015, pp. 344–352.

[171] Andrea Gasparetto et al. ‘A Survey on Text Classification Algorithms: From Text to Predictions’. In: *Information* 13.2 (Feb. 2022), p. 83. ISSN: 2078-2489.

[172] Andrea Gasparetto et al. ‘A survey on text classification: Practical perspectives on the Italian language’. In: *PLOS ONE* 17.7 (July 2022), pp. 1–46. ISSN: 1932-6203.

[173] Andrea Gasparetto et al. ‘Cross-Dataset Data Augmentation for Convolutional Neural Networks Training’. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. Vol. abs 1409 1556. IEEE Computer Society, Aug. 2018, pp. 910–915.

[174] Andrea Gasparetto et al. ‘Spatial Maps: From Low Rank Spectral to Sparse Spatial Functional Representations’. In: *2017 International Conference on 3D Vision (3DV)*. Vol. 33. IEEE Computer Society, Oct. 2017, pp. 477–485.

[175] P. V. Gehler et al. ‘Bayesian color constancy revisited’. In: *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, June 2008, pp. 1–8.

[176] Ghozlen, N B and Cerovic, Z G and Germain, C and Toutain, S and Latouche, G. ‘Non-destructive optical monitoring of grape maturation by proximal sensing’. In: *Sensors* 10.11 (Nov. 2010), pp. 10040–10068. ISSN: 1424-8220.

[177] A. Gijsenij, T. Gevers and J. van de Weijer. ‘Computational Color Constancy: Survey and Experiments’. In: *IEEE Transactions on Image Processing* 20.9 (Sept. 2011), pp. 2475–2489. ISSN: 1057-7149.

[178] Arjan Gijsenij, T. Gevers and Joost Weijer. ‘Generalized Gamut Mapping using Image Derivative Structures for Color Constancy’. In: *International Journal of Computer Vision* 86 (Jan. 2010), pp. 127–139. ISSN: 0920-5691.

[179] Jacob Gildenblat and contributors. *PyTorch library for CAM methods*. `https://github.com/jacobgil/pytorch-grad-cam`. 2021.

[180] Justin Gilmer et al. ‘Neural Message Passing for Quantum Chemistry’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. PMLR, Aug. 2017, pp. 1263–1272.

[181] Leilani H. Gilpin et al. ‘Explaining explanations: An overview of interpretability of machine learning’. In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. Ed. by Francesco Bonchi et al. IEEE. IEEE, Oct. 2018, pp. 80–89.

[182] Eleonora Giunchiglia and Thomas Lukasiewicz. ‘Coherent Hierarchical Multi-Label Classification Networks’. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., Oct. 2020, pp. 9662–9673.

[183] Keith Goatman et al. ‘Colour normalisation of retinal images’. In: (Jan. 2003).

[184] Jeremy Goecks et al. ‘How Machine Learning Will Transform Biomedicine’. en. In: *Cell* 181.1 (May 2020), pp. 92–101. ISSN: 0092-8674.

[185] Goel, Nidhi and Sehgal, Priti. ‘Fuzzy classification of pre-harvest tomatoes for ripeness estimation – An approach based on automatic rule learning using decision tree’. In: *Applied Soft Computing* 36 (Nov. 2015). Publisher: Elsevier BV, pp. 45–56. ISSN: 1568-4946.

[186] Gomez, A H and He, Y and Pereira, A G. ‘Non-destructive measurement of acidity, soluble solids and firmness of satsuma mandarin using vis/nir spectroscopy techniques’. In: *Journal of Food Engineering* 77 (Nov. 2006). Publisher: Progress on Bioproducts Processing and Food Safety, pp. 313–319. ISSN: 0260-8774.

[187] Jibing Gong et al. ‘Hierarchical Graph Transformer-Based Deep Learning Model for Large-Scale Multi-Label Text Classification’. In: *IEEE Access* 8 (Feb. 2020), pp. 30885–30896. ISSN: 2169-3536.

[188] José Ángel González et al. ‘Attentional Extractive Summarization’. In: *Applied Sciences* 13.3 (Jan. 2023), p. 1458. ISSN: 2076-3417.

[189] Ian Goodfellow, Yoshua Bengio and Aaron Courville. *Deep Learning*. MIT Press, Nov. 2016.

[190] Bryce Goodman and Seth Flaxman. 'European Union regulations on algorithmic decision making and a "right to explanation"'. en. In: *AI Mag.* 38.3 (Sept. 2017), pp. 50–57. ISSN: 0738-4602.

[191] Y. Gorishniy et al. 'Revisiting deep learning models for tabular data'. In: *Advances in Neural Information Processing Systems*. Ed. by Marc'Aurelio Ranzato et al. Vol. 34. Cornell University, June 2021, pp. 18932–18943.

[192] Yury Gorishniy, Ivan Rubachev and Artem Babenko. 'On embeddings for numerical features in tabular deep learning'. In: *Advances in Neural Information Processing Systems* 35 (Mar. 2022). Ed. by Sanmi Koyejo et al., pp. 24991–25004.

[193] Daniel J Gould et al. 'Patients' Views on AI for Risk Prediction in Shared Decision-Making for Knee Replacement Surgery: Qualitative Interview Study'. en. In: *J. Med. Internet Res.* 25 (Sept. 2023), e43632. ISSN: 1438-8871.

[194] Gowda, I N D and Huddar, A G. 'Studies on ripening changes in mango (Mangifera indica L.) fruits'. In: *Journal of Food Science and Technology-Mysore*. Vol. 38. 2. Association of Food Scientists and Technologists of India, Mar. 2001, pp. 135–137.

[195] Yash Goyal et al. *Counterfactual Visual Explanations*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Apr. 2019.

[196] Alex Graves. 'Generating Sequences With Recurrent Neural Networks'. In: *arXiv (preprint)* abs/1308.0850 (Aug. 2013). arXiv: 1308.0850 [cs.NE].

[197] Mara Graziani et al. 'A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences'. en. In: *Artif Intell Rev* 56.4 (May 2023), pp. 3473–3504. ISSN: 0269-2821.

[198] Gustavo Grieco, Matteo Maffei and Christoph Schneidewind. 'Echidna: A Fast Smart Contract Fuzzer'. In: *2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW)*. IEEE, 2020, pp. 435–444.

[199] Léo Grinsztajn, Edouard Oyallon and Gaël Varoquaux. 'Why do tree-based models still outperform deep learning on typical tabular data?' In: *Advances in neural information processing systems* 35 (Nov. 2022). Ed. by Sanmi Koyejo et al., pp. 507–520.

[200] Riccardo Guidotti et al. 'A Survey of Methods for Explaining Black Box Models'. In: *ACM Comput. Surv.* 51.5 (Aug. 2018), pp. 1–42. ISSN: 0360-0300.

[201] David Gunning et al. 'XAI–Explainable artificial intelligence'. In: *Science robotics* 4.37 (Dec. 2019), eaay7120. ISSN: 2470-9476.

[202] Longtao Guo et al. 'Reentrancy vulnerability detection based on graph convolutional networks and expert patterns under subspace mapping'. In: *Computers & Security* 142 (July 2024), p. 103894. ISSN: 0167-4048.

[203] Lyuying Guo et al. 'Syntax-Aware Retrieval Augmented Code Generation'. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Dec. 2023, pp. 9077–9089.

[204] Meng-Hao Guo et al. *Attention Mechanisms in Computer Vision: A Survey*. Nov. 2021.

[205] Holger A Haenssle et al. 'Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists'. In: *Annals of oncology* 29.8 (Aug. 2018), pp. 1836–1842. ISSN: 0923-7534.

[206] Halstead, Michael and McCool, Christopher and Denman, Simon and Perez, Tristan and Fookes, Clinton. 'Fruit Quantity and Ripeness Estimation Using a Robotic Vision System'. In: *IEEE Robotics and Automation Letters* 3.4 (Oct. 2018). Conference Name: IEEE Robotics and Automation Letters, pp. 2995–3002. ISSN: 2377-3766.

[207] William L. Hamilton, Rex Ying and Jure Leskovec. 'Inductive representation learning on large graphs'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Ed. by Isabelle Guyon et al. NIPS'17. Long Beach, California, USA: Curran Associates Inc., June 2017, pp. 1025–1035. ISBN: 9781510860964.

[208] Han, Kai and Wang, Yunhe and Chen, Hanting and Chen, Xinghao and Guo, Jianyuan and Liu, Zhenhua and Tang, Yehui and Xiao, An and Xu, Chunjing and Xu, Yixing and Yang, Zhaohui and Zhang, Yiman and Tao, Dacheng. 'A Survey on Vision Transformer'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45.1 (Jan. 2023). Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence, pp. 87–110. ISSN: 1939-3539.

[209] John T Hancock and Taghi M Khoshgoftaar. 'Survey on categorical data for neural networks'. In: *Journal of big data* 7.1 (May 2020), p. 28. ISSN: 2196-1115.

[210] Hartmann, Anja and Czauderna, Tobias and Hoffmann, Roberto and Stein, Nils and Schreiber, Falk. 'HTPheno: an image analysis pipeline for high-throughput plant phenotyping'. In: *BMC Bioinformatics* 12.1 (May 2011), p. 148. ISSN: 1471-2105.

[211] Nima Shiri Harzevili et al. 'A Systematic Literature Review on Automated Software Vulnerability Detection Using Machine Learning'. In: *ACM Comput. Surv.* 57.3 (Mar. 2025), pp. 1–36. ISSN: 0360-0300.

[212] Peter Hase and Mohit Bansal. *Evaluating Explainable AI: Which Algorithmic Explanations Help Users Predict Model Behavior?* Ed. by Dan Jurafsky et al. May 2020.

[213] Hazir, M H M and Shariff, A R M and Amiruddin, M D and Ramli, A R and Saripan, M Iqbal. 'Oil palm bunch ripeness classification using fluorescence technique'. In: *Journal of Food Engineering* 113.4 (Dec. 2012), pp. 534–540. ISSN: 0260-8774.

[214] Kaiming He et al. 'Deep residual learning for image recognition'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, June 2016, pp. 770–778.

[215] He, Kaiming and Gkioxari, Georgia and Dollar, Piotr and Girshick, Ross. 'Mask R-CNN'. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. event-place: Venice. IEEE, Oct. 2017, pp. 2980–2988.

[216] Stefan Hegselmann et al. *TabLLM: Few-shot Classification of Tabular Data with Large Language Models*. Oct. 2022.

[217] Herman, H and Susanto, A and Cenggoro, T W and Suharjito, S and Pardamean, B. 'Oil palm fruit image ripeness classification with computer vision using deep learning and visual attention'. In: *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)* 12.2 (June 2020), pp. 21–27. ISSN: 2180-1843.

[218] Liam Hiley, Alun D. Preece and Yulia Hicks. 'Explainable Deep Learning for Video Recognition Tasks: A Framework & Recommendations'. In: *CoRR* abs/1909.05667 (Sept. 2019). arXiv: 1909.05667.

[219] Geoffrey E Hinton, Sara Sabour and Nicholas Frosst. 'Matrix capsules with EM routing'. In: *International Conference on Learning Representations*. OpenReview.net, Feb. 2018.

[220] Geoffrey E. Hinton, Alex Krizhevsky and Sida D. Wang. 'Transforming Auto-Encoders'. In: *Artificial Neural Networks and Machine Learning – ICANN 2011*. Ed. by Timo Honkela et al. Vol. 6791. Berlin, Heidelberg: Springer Berlin Heidelberg, June 2011, pp. 44–51. ISBN: 978-3-642-21735-7.

[221] Tin Kam Ho. 'Random decision forests'. In: *Proceedings of 3rd International Conference on Document Analysis and Recognition*. Vol. 1. 1995, pp. 278–282.

[222] Hobson, Graeme E and Adams, Peter and Dixon, Timothy J. 'Assessing the colour of tomato fruit during ripening'. In: *Journal of the Science of Food and Agriculture* 34.3 (Mar. 1983). Publisher: Wiley, pp. 286–292. ISSN: 1097-0010.

[223] Sepp Hochreiter and Jürgen Schmidhuber. 'Long Short-term Memory'. In: *Neural computation* 9.8 (Dec. 1997), pp. 1735–1780. ISSN: 0899-7667.

[224] Robert Hoffman et al. 'Explaining Explanation, Part 4: A Deep Dive on Deep Nets'. In: *IEEE Intell. Syst.* 33.3 (May 2018), pp. 87–95. ISSN: 1541-1672.

[225] Robert R Hoffman, William J Clancy and Shane T Mueller. 'Explaining AI as an exploratory process: The peircean abduction model'. In: *arXiv (Cornell University)* abs/2009.14795 (Sept. 2020). ISSN: 2331-8422. arXiv: 2009.14795 [cs.AI].

[226] Robert R Hoffman, Timothy Miller and William J Clancey. 'Psychology and AI at a crossroads: How might complex systems explain themselves?' en. In: *Am. J. Psychol.* 135.4 (Dec. 2022), pp. 365–378. ISSN: 0002-9556.

[227] Fred Hohman et al. 'Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers'. In: *IEEE Transactions on Visualization and Computer Graphics* 25.8 (Aug. 2019). Conference Name: IEEE Transactions on Visualization and Computer Graphics, pp. 2674–2693. ISSN: 1941-0506.

[228] Andreas Holzinger et al. 'Explainable AI in Healthcare'. In: *BMJ* (2019).

[229] Andreas Holzinger et al. 'Causability and explainability of artificial intelligence in medicine'. en. In: *WIREs Data Mining and Knowledge Discovery* 9.4 (May 2019). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1312, e1312. ISSN: 1942-4795. (Visited on 29/05/2024).

[230] Andreas Holzinger et al. 'xxAI - Beyond Explainable Artificial Intelligence'. In: *xxAI - Beyond Explainable AI*. Springer International Publishing, 2022, pp. 3–10. ISBN: 9783031040832.

[231] Honkavaara, E and Kaivosoja, J and Makynen, J and Pellikka, I and Pesonen, L and Saari, H and Salo, H and Hakala, T and Marklelin, L and Rosnell, T. 'Hyperspectral reflectance signatures and point clouds for precision agriculture by light weight uav imaging system'. In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences I-7*. Vol. I-7. Copernicus GmbH, July 2012, pp. 353–358.

[232] Sara Hooker. 'Moving beyond "algorithmic bias is a data problem"'. In: *Patterns* 2.4 (Apr. 2021), p. 100241. ISSN: 2666-3899.

[233] Horea Mureşan and Mihai Oltean. *Fruit 360 dataset*. 2018.

[234] Horea Mureşan and Mihai Oltean. 'Fruit recognition from images using deep learning'. In: *Acta Universitatis Sapientiae, Informatica* 10.1 (Aug. 2018), pp. 26–42. ISSN: 1844-6086.

[235] Bo-Jian Hou and Zhi-Hua Zhou. 'Learning With Interpretable Structure From Gated RNN'. en. In: *IEEE Trans Neural Netw Learn Syst* 31.7 (July 2020), pp. 2267–2279. ISSN: 2162-237X.

[236] Hou, Saihui and Feng, Yushan and Wang, Zilei. 'VegFru: A Domain-Specific Dataset for Fine-Grained Visual Categorization'. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, pp. 541–549.

[237] Lijie Hu et al. *Editable Concept Bottleneck Models*. May 2024.

[238] Sihao Hu et al. 'Large Language Model-Powered Smart Contract Vulnerability Detection: New Perspectives'. In: *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*. Los Alamitos, CA, USA: IEEE Computer Society, Nov. 2023, pp. 297–306.

[239] Y. Hu, B. Wang and S. Lin. '$FC^4$: Fully Convolutional Color Constancy with Confidence-Weighted Pooling'. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Issn: 1063-6919. IEEE Computer Society, July 2017, pp. 330–339.

[240] Lei Huang et al. 'A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions'. In: *ACM Trans. Inf. Syst.* 43.2 (Mar. 2025), pp. 1–55. ISSN: 1046-8188.

[241] Wei Huang et al. 'Exploring Label Hierarchy in a Generative Way for Hierarchical Text Classification'. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 1116–1127.

[242] Wei Huang et al. 'Hierarchical Multi-label Text Classification: An Attention-based Recurrent Network Approach'. In: *Proceedings of the 28th ACM International Conference on Information and Knowledge Management, CIKM 2019, Beijing, China, November 3-7, 2019*. Ed. by Wenwu Zhu et al. ACM, Nov. 2019, pp. 1051–1060.

[243] X. Huang et al. ‘TabTransformer: Tabular data modeling using contextual embeddings’. In: *arXiv preprint arXiv:2012.06678* abs/2012.06678 (Dec. 2020). ISSN: 2331-8422.

[244] Yingren Huang et al. ‘Hierarchical multi-attention networks for document classification’. In: *International Journal of Machine Learning and Cybernetics* 12.6 (June 2021), pp. 1639–1647. ISSN: 1868-808X.

[245] Yue Huang et al. ‘Position: TRUSTLLM: trustworthiness in large language models’. In: *Proceedings of the 41st International Conference on Machine Learning*. ICML’24. Vienna, Austria: JMLR.org, 2024.

[246] Huang, Ching-Hsuan and He, Jiayang and Austin, Elena and Seto, Edmund and Novosselov, Igor. ‘Assessing the value of complex refractive index and particle density for calibration of low-cost particle matter sensor for size-resolved particle count and PM2.5 measurements’. In: *PLoS One* 16.11 (Nov. 2021). Publisher: Public Library of Science, e0259745. ISSN: 1932-6203.

[247] Sara Bronwen Hunter, Fiona Mathews and Julie Weeds. ‘Using hierarchical text classification to investigate the utility of machine learning in automating online analyses of wildlife exploitation’. In: *Ecological Informatics* 75 (July 2023), p. 102076. ISSN: 1574-9541.

[248] Hussain, Israr and Wu, Wei Long and Hua, He Qian and Hussain, Naeem. ‘Intra-Class Recognition of Fruits Using DCNN for Commercial Trace Back-System’. In: *Proceedings of the 2019 4th International Conference on Multimedia Systems and Signal Processing*. Vol. 148. ICMSSP ’19. Guangzhou, China: Association for Computing Machinery, May 2019, pp. 194–199. ISBN: 9781450371711.

[249] Forrest N. Iandola et al. ‘SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <1MB model size’. In: *CoRR* abs/1602.07360 (2016). arXiv: 1602.07360.

[250] Alexey Ignatiev. ‘Towards Trustable Explainable AI’. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Ed. by Christian Bessiere. Yokohama, Japan: International Joint Conferences on Artificial Intelligence Organization, July 2020, pp. 5154–5158.

[251] SangHun Im et al. ‘Hierarchical Text Classification as Sub-hierarchy Sequence Generation’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 37.11 (June 2023). Ed. by Brian Williams, Yiling Chen and Jennifer Neville, pp. 12933–12941. ISSN: 2159-5399.

[252] ‘Interpretable and Explainable Machine Learning Methods: A Systematic Literature Review’. In: *arXiv preprint arXiv:2312.17584* (2023).

[253] Israr Hussain and Qianhua He and Zhuliang Chen and Wei Xie. *Fruit Recognition dataset*. July 2018.

[254] Alon Jacovi and Yoav Goldberg. ‘Aligning Faithful Interpretations with their Social Attribution’. en. In: *Transactions of the Association for Computational Linguistics* 9 (Mar. 2021), pp. 294–310. ISSN: 2307-387X. eprint: https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00367/1923972/tacl_a_00367.pdf.

[255] Alon Jacovi and Yoav Goldberg. ‘Towards Faithfully Interpretable NLP Systems: How Should We Define and Evaluate Faithfulness?’ In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4198–4205.

[256] Sarthak Jain and Byron C. Wallace. ‘Attention is not Explanation’. en. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Vol. 1. Long and Short Papers. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 3543–3556. (Visited on 31/05/2021).

[257] Sarthak Jain et al. ‘Learning to Faithfully Rationalize by Construction’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, July 2020, pp. 4459–4473.

[258] Vikas Kumar Jain and Meenakshi Tripathi. ‘An integrated deep learning model for Ethereum smart contract vulnerability detection’. In: *International Journal of Information Security* 23.1 (Feb. 2024), pp. 557–575. ISSN: 1615-5262.

[259] Hyeju Jang et al. ‘Classification of Alzheimer’s Disease Leveraging Multi-task Machine Learning Analysis of Speech and Eye-Movement Data’. In: *Frontiers in Human Neuroscience* 15 (Sept. 2021), p. 716670. ISSN: 1662-5161.

[260] Jaradat, A A and Zaid, A. ‘Quality traits of date palm fruits in a center of origin and center of diversity’. In: *International Journal of Food, Agriculture and Environment* 2 (2004), pp. 208–217.

[261] Ganesh Jawahar, Benoît Sagot and Djamé Seddah. ‘What Does BERT Learn about the Structure of Language?’ In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David R. Traum and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 3651–3657.

[262] Jha, S N and Chopra, Sangeeta and Kingsly, A R P. ‘Modeling of color values for nondestructive evaluation of maturity of mango’. In: *Journal of Food Engineering* 78.1 (Jan. 2007). Publisher: Elsevier BV, pp. 22–26. ISSN: 0260-8774.

[263] Jha, S N and Kingsly, A R P and Chopra, S. ‘Non-destructive Determination of Firmness and Yellowness of Mango during Growth and Storage using Visual Spectroscopy’. In: *Biosystems Engineering* 94.3 (July 2006). Publisher: Elsevier BV, pp. 397–402. ISSN: 1537-5110.

[264] Ke Ji et al. ‘Hierarchical Verbalizer for Few-Shot Hierarchical Text Classification’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 2918–2933.

[265] Feng Jiang et al. ‘The Role of AI in Healthcare Collaborative Decision Making’. In: *IEEE Trans. Neural Netw. Learn. Syst.* (2017).

[266] Hang Jiang et al. ‘Financial News Annotation by Weakly-Supervised Hierarchical Multi-label Learning’. In: *Proceedings of the Second Workshop on Financial Technology and Natural Language Processing*. Kyoto, Japan: -, Jan. 2020, pp. 1–7.

[267] Ting Jiang et al. ‘Exploiting Global and Local Hierarchies for Hierarchical Text Classification’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva and Yue Zhang. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 4030–4039.

[268] Yufan Jiang et al. ‘Improved Differentiable Architecture Search for Language Modeling and Named Entity Recognition’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3585–3590.

[269] Karen Spärck Jones. ‘A statistical interpretation of term specificity and its application in retrieval’. In: *J. Doc.* 28.1 (1972), pp. 11–21. ISSN: 0022-0418.

[270] Armand Joulin et al. ‘Bag of Tricks for Efficient Text Classification’. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*. Ed. by Mirella Lapata, Phil Blunsom and Alexander Koller. Valencia, Spain: Association for Computational Linguistics, May 2017, pp. 427–431.

[271] Juan Jovel and Russell Greiner. ‘An Introduction to Machine Learning Approaches for Biomedical Research’. en. In: *Front. Med.* 8 (Dec. 2021), p. 771607. ISSN: 2296-858X.

[272] Daniel Jurafsky and James Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd (draft). Upper Saddle River, NJ, USA: Prentice-Hall, Inc., Dec. 2020, pp. 30–35.

[273] Daniel Kahneman. ‘Fast and slow thinking’. In: *Allen Lane and Penguin Books, New York* (2011).

[274] Jiyeon Kang et al. *Score-informed Neural Operator for Enhancing Ordering-based Causal Discovery*. Aug. 2025.

[275] Jared Kaplan et al. ‘Scaling Laws for Neural Language Models’. In: *arXiv (preprint)* abs/2001.08361 (Jan. 2020). ISSN: 2331-8422.

[276] Amir-Hossein Karimi et al. ‘A Survey of Algorithmic Recourse: Contrastive Explanations and Consequential Recommendations’. In: *ACM Comput. Surv.* 55.5 (Dec. 2022), pp. 1–29. ISSN: 0360-0300.

[277] Amir-Hossein Karimi et al. *Model-Agnostic Counterfactual Explanations for Consequential Decisions*. May 2019.

[278] Liran Katzir, Gal Elidan and Ran El-Yaniv. ‘Net-dnf: Effective deep modeling of tabular data’. In: *International conference on learning representations*. 2020.

[279] Kawano, Sumio and Abe, Hideyuki and Iwamoto, Mutsuo. ‘Development of a Calibration Equation with Temperature Compensation for Determining the Brix Value in Intact Peaches’. In: *Journal of Near Infrared Spectroscopy* 3.4 (Oct. 1995). Publisher: SAGE Publishing, pp. 211–218. ISSN: 1751-6552. (Visited on 21/02/2023).

[280] Rémi Kazmierczak et al. ‘Explainability and vision foundation models: A survey’. In: *Inf. Fusion* 122.C (June 2025), p. 103184. ISSN: 1566-2535.

[281] G. Ke et al. ‘DeepGBM: A deep learning framework distilled by GBDT for online prediction tasks’. In: *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Ed. by Ankur Teredesai et al. ACM, July 2019, pp. 384–394.

[282] Jun Kevin and Pujianto Yugopuspito. ‘SmartLLM: Smart Contract Auditing using Custom Generative AI’. In: *2025 International Conference on Computer Sciences, Engineering, and Technology Innovation (ICoCSETI)*. IEEE, Jan. 2025, pp. 260–265.

[283] Ashkan Khakzar et al. ‘Do Explanations Explain? Model Knows Best’. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2022, pp. 10234–10243.

[284] Salman Khan et al. ‘Transformers in Vision: A Survey’. In: *ACM Comput. Surv.* 54.10s (Sept. 2022), pp. 1–41. ISSN: 0360-0300.

[285] Khandelwal, U and Clark, K and Jurafsky, D and Kaiser, L. *Sample efficient text summarization using a single Pre-Trained transformer*. May 2019.

[286] Been Kim et al. ‘Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)’. In: *arXiv (Cornell University)* (Nov. 2017).

[287] Eun-Sol Kim et al. ‘Temporal Attention Mechanism with Conditional Inference for Large-Scale Multi-label Video Classification’. en. In: *Computer Vision – ECCV 2018 Workshops*. Ed. by Laura Leal-Taixé and Stefan Roth. Vol. 11132. Springer, Cham, Sept. 2018, pp. 306–316. (Visited on 31/05/2021).

[288] Hyo-Eun Kim et al. ‘Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study’. In: *The Lancet Digital Health* 2.3 (Mar. 2020), e138–e148. ISSN: 2589-7500.

[289] Jenia Kim, Henry Maathuis and Danielle Sent. ‘Human-centered evaluation of explainable AI applications: a systematic review’. In: *Frontiers in Artificial Intelligence* 7 (2024). DOI: 10.3389/frai.2024.1456486.

[290] Yoon Kim. ‘Convolutional Neural Networks for Sentence Classification’. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. Association for Computational Linguistics, Oct. 2014, pp. 1746–1751.

[291] Pieter-Jan Kindermans et al. ‘The (Un)reliability of Saliency Methods’. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Vol. 11700. Cham: Springer International Publishing, 2019, pp. 267–280. ISBN: 978-3-030-28954-6.

[292] Diederik P Kingma and Jimmy Ba. ‘Adam: A method for stochastic optimization’. In: *arXiv preprint arXiv:1412.6980* (Dec. 2014).

[293] Svetlana Kiritchenko et al. ‘Learning and Evaluation in the Presence of Class Hierarchies: Application to Text Categorization’. In: *Advances in Artificial Intelligence*. Ed. by Luc Lamontagne and Mario Marchand. Vol. 4013. Berlin, Heidelberg: Springer Berlin Heidelberg, June 2006, pp. 395–406. ISBN: 978-3-540-34630-2.

[294] Klasson, Marcus and Zhang, Cheng and Kjellström, Hedvig. ‘A Hierarchical Grocery Store Image Dataset With Visual and Semantic Labels’. In: *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2019, pp. 491–500.

[295] Bryan Klimt and Yiming Yang. ‘The Enron Corpus: A New Dataset for Email Classification Research’. In: *Proceedings of the 15th European Conference on Machine Learning*. Ed. by Jean-François Boulicaut et al. Vol. 3201. ECML’04. Pisa, Italy: Springer-Verlag, Sept. 2004, pp. 217–226. ISBN: 3540231056.

[296] Kodors, Sergejs and Lacis, Gunars and Sokolova, Olga and Zhukov, Vitaliy and Apeinans, Ilmars and Bartulsons, Toms. ‘Apple scab detection using CNN and Transfer Learning’. In: *Agronomy Research* 19 (Apr. 2021), pp. 507–519.

[297] Kodors, Sergejs and Lacis, Gunars and Sokolova, Olga and Zhukov, Vitaliy and Apeinans, Ilmars and Bartulsons, Toms. *Apples infected by scab dataset*. 2021.

[298] Kodors, Sergejs and Lacis, Gunars and Sokolova, Olga and Zhukov, Vitaliy and Apeinans, Ilmars and Bartulsons, Toms. *Leaves of apples infected by scab dataset*. 2021.

[299] Narine Kokhlikyan et al. ‘Captum: A unified and generic model interpretability library for pytorch’. In: *arXiv preprint arXiv:2009.07896* abs/2009.07896 (Sept. 2020). ISSN: 2331-8422.

[300] Koklu, Murat and Kursun, Ramazan and Taspinar, Yavuz Selim and Cinar, Ilkay. ‘Classification of Date Fruits into Genetic Varieties Using Image Analysis’. In: *Mathematical Problems in Engineering* 2021 (Nov. 2021), p. 4793293. ISSN: 1024-123X.

[301] Koklu, Murat and Kursun, Ramazan and Taspinar, Yavuz Selim and Cinar, Ilkay. *Date Image Dataset*. 2021.

[302] Daphne Koller and Mehran Sahami. ‘Hierarchically Classifying Documents Using Very Few Words’. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. Ed. by Douglas H. Fisher. ICML ’97. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., July 1997, pp. 170–178. ISBN: 1558604863.

[303] Aneesh Komanduri, Karuna Bhaila and Xintao Wu. *CausalVLBench: Benchmarking Visual Causal Reasoning in Large Vision-Language Models*. May 2025.

[304] Aris Kosmopoulos et al. ‘Evaluation measures for hierarchical classification: a unified view and novel approaches’. In: *Data Mining and Knowledge Discovery* 29.3 (May 2015), pp. 820–865. ISSN: 1573-756X.

[305] Georgios Kostopoulos, Gregory Davrazos and Sotiris Kotsiantis. ‘Explainable Artificial Intelligence-Based Decision Support Systems: A Recent Review’. In: *Electronics* 13.14 (July 2024), p. 2842. ISSN: 2079-9292.

[306] Kamran Kowsari et al. ‘HDLTex: Hierarchical Deep Learning for Text Classification’. In: *2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. Ed. by Xuewen Chen et al. Cancun, Mexico: IEEE, Dec. 2017, pp. 364–371.

[307] Kamran Kowsari et al. ‘Text Classification Algorithms: A Survey’. In: *Information.* 10.4 (May 2019), p. 150. ISSN: 2078-2489.

[308] Kamran Kowsari et al. *Web of Science Dataset*. Mar. 2018.

[309] Mark A. Kramer. ‘Nonlinear principal component analysis using autoassociative neural networks’. In: *AIChE Journal* 37.2 (Feb. 1991), pp. 233–243. ISSN: 0001-1541.

[310] Milan Krendzelak and Frantisek Jakab. ‘Hierarchical Text Classification Using CNNS with Local Approaches’. In: *Comput. Informatics* 39.5 (2020), pp. 907–924. ISSN: 1335-9150.

[311] J. von Kries. *Chromatic adaption*. 1902.

[312] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. ‘ImageNet Classification with Deep Convolutional Neural Networks’. In: *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*. Ed. by Peter L. Bartlett et al. Vol. 25. Nips’12. Lake Tahoe, Nevada: Curran Associates Inc., Dec. 2012, pp. 1097–1105.

[313] Taku Kudo and John Richardson. ‘SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing’. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Eduardo Blanco and Wei Lu. Brussels, Belgium: Association for Computational Linguistics, Nov. 2018, pp. 66–71.

[314] I. Elizabeth Kumar et al. ‘Problems with Shapley-value-based explanations as feature importance measures’. In: *Proceedings of the 37th International Conference on Machine Learning*. Vol. 119. Proceedings of Machine Learning Research. PMLR, July 2020, pp. 5491–5500.

[315] Yogesh Kumar et al. ‘Artificial intelligence in disease diagnosis: a systematic literature review, synthesizing framework and future research agenda’. en. In: *J. Ambient Intell. Humaniz. Comput.* 14.7 (July 2023), pp. 8459–8486. ISSN: 1868-5137.

[316] Cao Yu-kun et al. ‘Hierarchical Label Text Classification Method with Deep-Level Label-Assisted Classification’. In: *2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS)*. IEEE, May 2023, pp. 1467–1474.

[317] Shinjini Kundu. ‘AI in medicine must be explainable’. en. In: *Nat. Med.* 27.8 (Aug. 2021), p. 1328. ISSN: 1078-8956.

[318] Jingun Kwon et al. ‘Hierarchical Label Generation for Text Classification’. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 625–632.

[319] Cosimo Laneve et al. ‘Assessing Code Understanding in LLMs’. In: *Formal Techniques for Distributed Objects, Components, and Systems - 45th IFIP WG 6.1 International Conference, FORTE 2025*. Ed. by Carla Ferreira and Claudio Antares Mezzina. Vol. 15732. The provided DOI and ISBN could not be resolved. The entry is based on an arXiv preprint for a future conference. Springer Science+Business Media, Mar. 2025, pp. 202–210. ISBN: 978-3031954962.

[320] Ken Lang. ‘NewsWeeder: Learning to Filter Netnews’. In: *Machine Learning Proceedings 1995*. Ed. by Armand Prieditis and Stuart Russell. San Francisco (CA): Morgan Kaufmann, July 1995, pp. 331–339. ISBN: 978-1-55860-377-6.

[321] Sebastian Lapuschkin et al. ‘Unmasking Clever Hans predictors and assessing what machines really learn’. In: *Nature Communications* 10.1 (Mar. 2019), p. 1096. ISSN: 2041-1723.

[322] Colin Lea et al. ‘Temporal Convolutional Networks for Action Segmentation and Detection’. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, July 2017, pp. 1003–1012.

[323] Y. LeCun et al. ‘Backpropagation Applied to Handwritten Zip Code Recognition’. In: *Neural Computation* 1.4 (Dec. 1989), pp. 541–551. ISSN: 0899-7667.

[324] Tao Lei. ‘Interpretable neural models for natural language processing’. eng. Accepted: 2017-05-11T19:59:27Z. Thesis. Massachusetts Institute of Technology, 2017. (Visited on 31/05/2021).

[325] Dmitry Lepikhin et al. ‘GShard: Scaling Giant Models with Conditional Computation and Automatic Sharding’. In: *International Conference on Learning Representations (ICLR 2021)*. Vol. abs/2006.16668. Vienna, Austria: Cornell University, June 2021.

[326] Arnon Levy. ‘Carl F. Craver, Explaining what? Review of explaining the brain: mechanisms and the mosaic unity of neuroscience: Clarendon Press–Oxford University Press, 2007, 272 pp, $49.50 (06)’. In: *Biology & Philosophy* 24.1 (Aug. 2008), pp. 137–145. ISSN: 1572-8404.

[327] David D. Lewis et al. ‘RCV1: A New Benchmark Collection for Text Categorization Research’. In: *J. Mach. Learn. Res.* 5 (Dec. 2004), pp. 361–397. ISSN: 1532-4435.

[328] Patrick Lewis et al. ‘Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks’. In: *Advances in Neural Information Processing Systems*. Ed. by Hugo Larochelle et al. Vol. 33. University College London, May 2020, pp. 9459–9474.

[329] Bo Li et al. ‘Trustworthy AI: From Principles to Practices’. en. In: *ACM Comput. Surv.* 55.9 (Jan. 2023), pp. 1–46. ISSN: 0360-0300.

[330] Fei Li, Zhengyi Chen and Yanyan Wang. ‘HLC-KEPLM: Hierarchical Label Classification Based on Knowledge-Enhanced Pretrained Language Model for Chinese Telecom’. In: *2023 4th International Conference on Intelligent Computing and Human-Computer Interaction (ICHCI)*. IEEE, Aug. 2023, pp. 262–266.

[331] Hongjing Li et al. ‘Distinguishability Calibration to In-Context Learning’. In: *Findings of the Association for Computational Linguistics: EACL 2023*. Ed. by Andreas Vlachos and Isabelle Augenstein. Vol. abs/2302.06198. Dubrovnik, Croatia: Association for Computational Linguistics, May 2023, pp. 1385–1397.

[332] Jia Li et al. ‘Structured Chain-of-Thought Prompting for Code Generation’. In: *ACM Trans. Softw. Eng. Methodol.* 34.2 (Feb. 2025), pp. 1–23. ISSN: 1049-331X.

[333] Qian Li et al. ‘A Survey on Text Classification: From Traditional to Deep Learning’. In: *ACM Trans. Intell. Syst. Technol.* 13.2 (May 2022), pp. 1–41. ISSN: 2157-6904.

[334] Qimai Li, Zhichao Han and Xiao-Ming Wu. ‘Deeper Insights into Graph Convolutional Networks for Semi-Supervised Learning’. In: *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*. Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. Vol. 32. AAAI’18/IAAI’18/EAAI’18 1. New Orleans, Louisiana, USA: AAAI Press, May 2018, pp. 3538–3545. ISBN: 978-1-57735-800-8.

[335] Renxuan Albert Li et al. ‘Analysis of Hierarchical Multi-Content Text Classification Model on B-SHARP Dataset for Early Detection of Alzheimer’s Disease’. In: *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, AACL/IJCNLP 2020, Suzhou, China, December 4-7, 2020*. Ed. by Kam-Fai Wong, Kevin Knight and Hua Wu. Association for Computational Linguistics, 2020, pp. 358–365.

[336] Shuang Li et al. ‘Diversity Regularized Spatiotemporal Attention for Video-Based Person Re-Identification’. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, June 2018, pp. 369–378. (Visited on 31/05/2021).

[337] Wei Li et al. ‘Object detection based on an adaptive attention mechanism’. en. In: *Scientific Reports* 10.1 (July 2020). Number: 1 Publisher: Nature Publishing Group, p. 11307. ISSN: 2045-2322. (Visited on 01/06/2021).

[338] Xingyu Li, Karunesh Arora and Saman Alaniazar. ‘Mixed-Model Text Classification Framework Considering the Practical Constraints’. In: *2019 SECOND INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE FOR INDUSTRIES (AI4I 2019)*. 2nd IEEE International Conference on Artificial Intelligence for Industries (AI4I), Sep 25–27, 2019. IEEE; IEEE Comp Soc. Laguna Hills, CA, USA: IEEE, Sept. 2019, pp. 67–70. ISBN: 978-1-7281-4087-2.

[339] Yang Li et al. ‘Attention Based CNN-ConvLSTM for Pedestrian Attribute Recognition’. en. In: *Sensors* 20.3 (Jan. 2020). Number: 3 Publisher: Multidisciplinary Digital Publishing Institute, p. 811. ISSN: 1424-8220. (Visited on 31/05/2021).

[340] Z. Li, B. Liu and J. Tang. ‘Robust structured subspace learning for data representation’. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37.10 (Oct. 2015), pp. 2085–2098. ISSN: 0162-8828.

[341] Zhenyang Li et al. ‘VideoLSTM convolves, attends and flows for action recognition’. In: *Computer Vision and Image Understanding* 166.C (Jan. 2018), pp. 41–50. ISSN: 1077-3142. (Visited on 31/05/2021).

[342] Zongwei Li et al. ‘SCALM: Detecting Bad Practices in Smart Contracts Through LLMs’. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Ed. by Toby Walsh, Julie Shah and Zico Kolter. Vol. 39. 1. Association for the Advancement of Artificial Intelligence (AAAI), Feb. 2025, pp. 470–477.

[343] Li, Bo and Lecourt, Julien and Bishop, Gerard. ‘Advances in Non-Destructive Early Assessment of Fruit Ripeness towards Defining Optimal Time of Harvest and Yield Prediction–A Review’. In: *Plants* 7.1 (Mar. 2018). Number: 1 Publisher: Multidisciplinary Digital Publishing Institute, p. 3. ISSN: 2223-7747. (Visited on 21/02/2023).

[344] Li, Han and Lee, Won Suk and Wang, Ku. ‘Identifying blueberry fruit of different growth stages using natural outdoor color images’. In: *Computers and Electronics in Agriculture* 106 (Aug. 2014). Publisher: Elsevier BV, pp. 91–101. ISSN: 0168-1699.

[345] Li, M and Slaughter, D C and Thompson, J F. ‘Optical chlorophyll sensing system for banana ripening’. In: *Postharvest Biology and Technology* 12.3 (Dec. 1997), pp. 273–283. ISSN: 0925-5214.

[346] Chen Liang et al. ‘Visual Abductive Reasoning’. In: *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. New Orleans, LA, USA: IEEE, June 2022, pp. 15544–15554.

[347] Xin Liang et al. ‘F-HMTC: Detecting Financial Events for Investment Decisions Based on Neural Hierarchical Multi-Label Text Classification’. In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*. Ed. by Christian Bessiere. ijcai.org, July 2020, pp. 4490–4496.

[348] Brian Y Lim et al. ‘Diagrammatization: Rationalizing with diagrammatic AI explanations for abductive-deductive reasoning on hypotheses’. In: *arXiv (Cornell University)* abs/2302.01241 (Feb. 2023). ISSN: 2331-8422. arXiv: 2302.01241 [cs.AI].

[349] Liming, X and Yanchao, Z. ‘Automated strawberry grading system based on image processing’. In: *Computers and Electronics in Agriculture*. Special issue on computer and computing technologies in agriculture 71 (May 2010), S32–S39. ISSN: 0168-1699. (Visited on 21/02/2023).

[350] Jimmy Lin. ‘The Neural Hype and Comparisons Against Weak Baselines’. In: *ACM SIGIR Forum* 52.2 (Jan. 2019), pp. 40–51. ISSN: 0163-5840. (Visited on 18/06/2024).

[351] Pantelis Linardatos, Vasilis Papastefanopoulos and Sotiris Kotsiantis. ‘Explainable ai: A review of machine learning interpretability methods’. In: *Entropy* 23.1 (Dec. 2020), p. 18. ISSN: 1099-4300.

[352] Robert K. Lindsay et al. ‘DENDRAL: A case study of the first expert system for scientific hypothesis formation’. In: *Artificial Intelligence* 61.2 (June 1993), pp. 209–261. ISSN: 0004-3702.

[353] Zachary C Lipton. ‘The Mythos of Model Interpretability’. In: *Queueing Syst.* 16.3 (June 2016), pp. 36–43. ISSN: 1542-7730. arXiv: 1606.03490 [cs.LG].

[354] Hankai Liu, Xianying Huang and Xiaoyang Liu. ‘Improve label embedding quality through global sensitive GAT for hierarchical text classification’. In: *Expert Systems with Applications* 238 (Mar. 2024), p. 122267. ISSN: 0957-4174.

[355] Hui Liu et al. ‘Improving Pretrained Models for Zero-shot Multi-label Text Classification through Reinforced Label Hierarchy Reasoning’. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 1051–1062.

[356] Jingang Liu et al. ‘Hierarchical Comprehensive Context Modeling for Chinese Text Classification’. In: *IEEE Access* 7 (Oct. 2019), pp. 154546–154559. ISSN: 2169-3536.

[357] Jingzhou Liu et al. ‘Deep Learning for Extreme Multi-Label Text Classification’. In: *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ed. by Noriko Kando et al. SIGIR ’17. Association for Computing Machinery, Aug. 2017, pp. 115–124. ISBN: 9781450350228.

[358] Leibo Liu et al. ‘Automated ICD coding using extreme multi-label long text transformer-based models’. In: *Artificial Intelligence in Medicine* 144 (Oct. 2023), p. 102662. ISSN: 0933-3657.

[359] Liqun Liu et al. ‘NeuralClassifier: An Open-source Neural Hierarchical Multi-label Text Classification Toolkit’. In: *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28 - August 2, 2019, Volume 3: System Demonstrations*. Ed. by Marta R. Costa-jussà and Enrique Alfonseca. Association for Computational Linguistics, July 2019, pp. 87–92.

[360] Pengfei Liu, Xipeng Qiu and Xuanjing Huang. ‘Recurrent Neural Network for Text Classification with Multi-Task Learning’. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. Ed. by Subbarao Kambhampati. Vol. abs/1605.05101. IJCAI’16. New York, New York, USA: AAAI Press, May 2016, pp. 2873–2879. ISBN: 9781577357704.

[361] Xiao Liu et al. ‘P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks’. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 61–68.

[362] Xiaoxuan Liu et al. ‘Artificial Intelligence in Medicine: A Review of Current and Future Applications’. In: *J. Intern. Med.* (2020).

[363] Ye Liu et al. ‘Enhancing Hierarchical Text Classification through Knowledge Graph Integration’. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 5797–5810.

[364] Yinhan Liu et al. ‘RoBERTa: A Robustly Optimized BERT Pretraining Approach’. In: *arXiv* abs/1907.11692 (July 2019). ISSN: 2331-8422. arXiv: 1907.11692 [cs.CL].

[365] Liu, Changhong and Liu, Wei and Chen, Wei and Yang, Jianbo and Zheng, Lei. 'Feasibility in multispectral imaging for predicting the content of bioactive compounds in intact tomato fruit'. In: *Food Chemistry* 173 (May 2015), pp. 482–488. ISSN: 0308-8146. (Visited on 21/02/2023).

[366] Llobet, E and Hines, E L and Gardner, J W and Franco, S. 'Non-destructive banana ripeness determination using a neural network-based electronic nose'. In: *Measurement Science and Technology* 10.6 (June 1999), pp. 538–548. ISSN: 0957-0233.

[367] Luigi Lomasto et al. 'An Automatic Text Classification Method Based on Hierarchical Taxonomies, Neural Networks and Document Embedding: The NETHIC Tool'. In: *Enterprise Information Systems*. Ed. by Joaquim Filipe et al. Cham: Springer International Publishing, 2020, pp. 57–77. ISBN: 978-3-030-40783-4.

[368] Tania Lombrozo. 'The structure and function of explanations'. In: *Trends in Cognitive Sciences* 10.10 (Oct. 2006), pp. 464–470. ISSN: 1364-6613.

[369] Loomis, R S and Williams, W A. 'Maximum crop productivity: An estimate 1'. In: *Crop Science* 3.1 (Jan. 1963). Publisher: Wiley, pp. 67–72.

[370] Ilya Loshchilov and Frank Hutter. 'Decoupled Weight Decay Regularization'. In: *International Conference on Learning Representations*. OpenReview.net, 2019.

[371] Jörn Lötsch, Dario Kringel and Alfred Ultsch. 'Explainable Artificial Intelligence (XAI) in Biomedicine: Making AI Decisions Trustworthy for Physicians and Patients'. en. In: *BioMedInformatics* 2.1 (Dec. 2021), pp. 1–17. ISSN: 2673-7426.

[372] Zhiyong Lu. 'PubMed and beyond: a survey of web tools for searching biomedical literature'. In: *Database (Oxford)* 2011.0 (Jan. 2011), baq036. ISSN: 1758-0463.

[373] Lu, Renfu and Peng, Yankun. 'Hyperspectral Scattering for assessing Peach Fruit Firmness'. In: *Biosystems Engineering* 93.2 (Feb. 2006). Publisher: Elsevier BV, pp. 161–171. ISSN: 1537-5110.

[374] Scott Lundberg. 'A unified approach to interpreting model predictions'. In: *arXiv preprint arXiv:1705.07874*. NIPS'17 30 (May 2017). Ed. by Isabelle Guyon et al., pp. 4768–4777.

[375] Thang Luong, Hieu Pham and Christopher D. Manning. 'Effective Approaches to Attention-based Neural Machine Translation'. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Ed. by Lluís Màrquez et al. Vol. abs/1508.04025. Lisbon, Portugal: Association for Computational Linguistics, Sept. 2015, pp. 1412–1421.

[376] Lurie, S and Friedman, H and Weksler, A and Dagar, A and Zerbini, P Eccher. 'Maturity assessment at harvest and prediction of softening in an early and late season melting peach'. In: *Postharvest Biology and Technology* 76 (Feb. 2013), pp. 10–16. ISSN: 0925-5214.

[377] Loi Luu et al. 'Making Smart Contracts Smarter'. In: *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. Ed. by Edgar R. Weippl et al. ACM, Oct. 2016, pp. 254–269.

[378] Volodymyr Lyubinets, Taras Boiko and Deon Nicholas. 'Automated Labeling of Bugs and Tickets Using Attention-Based Mechanisms in Recurrent Neural Networks'. In: *2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP)*. Vol. 9. IEEE, Aug. 2018, pp. 271–275.

[379] Kefan Ma et al. 'LED: Label Correlation Enhanced Decoder for Multi-Label Text Classification'. In: *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, June 2023, pp. 1–5.

[380] Wei Ma et al. 'Combining Fine-tuning and LLM-based Agents for Intuitive Smart Contract Auditing with Justifications'. In: *2025 IEEE/ACM 47th International Conference on Software Engineering (ICSE)*. Los Alamitos, CA, USA: IEEE Computer Society, May 2025, pp. 330–342.

[381] Yinglong Ma, Jingpeng Zhao and Beihong Jin. 'A Hierarchical Fine-Tuning Approach Based on Joint Embedding of Words and Parent Categories for Hierarchical Multi-label Text Classification'. In: *Artificial Neural Networks and Machine Learning - ICANN 2020 - 29th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 15-18, 2020, Proceedings, Part II*. Ed. by Igor Farkas, Paolo Masulli and Stefan Wermter. Vol. 12397. Lecture Notes in Computer Science. Springer, May 2020, pp. 746–757. ISBN: 978-3030616151.

[382] Yinglong Ma et al. 'Hybrid embedding-based text representation for hierarchical multi-label text classification'. In: *Expert Syst. Appl.* 187 (Jan. 2022), p. 115905. ISSN: 0957-4174.

[383] Ma, Guang and Fu, Xia-Ping and Zhou, Ying and Ying, Yi-Bin and Xu, Hui-Rong and Xie, Li-Juan and Lin, Tao. 'Nondestructive sugar content determination of peaches by using near infrared spectroscopy technique'. In: *Guang Pu Xue Yu Guang Pu Fen Xi* 27.5 (May 2007), pp. 907–910.

[384] Maciej Adamiak. *Lemons quality control dataset*. July 2020.

[385] D. J. C. MacKay. 'Bayesian methods for backpropagation networks'. In: *Models of Neural Networks III*. Springer, 1994, pp. 211–254.

[386] Andrzej Maćkiewicz and Waldemar Ratajczak. 'Principal components analysis (PCA)'. In: *Computers & Geosciences* 19.3 (Mar. 1993), pp. 303–342. ISSN: 0098-3004.

[387] Andreas Madsen et al. 'Evaluating the Faithfulness of Importance Measures in NLP by Recursively Masking Allegedly Important Tokens and Retraining'. In: *arXiv (Cornell University)* (Oct. 2021), pp. 1731–1751. arXiv: 2110.08412 [cs.CL].

[388] Mahesh, S and Jayas, D S and Paliwal, J and White, N D G. 'Hyperspectral imaging to classify and monitor quality of agricultural materials'. In: *Journal of Stored Products Research* 61 (Mar. 2015). Publisher: Elsevier BV, pp. 17–26. ISSN: 0022-474X.

[389] Mahesh, S. and Jayas, D. S. and Paliwal, J. and White, N. D. G. ‘Comparison of Partial Least Squares Regression (PLSR) and Principal Components Regression (PCR) Methods for Protein and Hardness Predictions using the Near-Infrared (NIR) Hyperspectral Images of Bulk Samples of Canadian Wheat’. In: *Food and Bioprocess Technology* 8.1 (Jan. 2015), pp. 31–40. ISSN: 1935-5149. (Visited on 21/02/2023).

[390] Makky, Muhammad and Soni, Peeyush. ‘In situ quality assessment of intact oil palm fresh fruit bunches using rapid portable non-contact and non-destructive approach’. In: *Journal of Food Engineering* 120 (Jan. 2014). Publisher: Elsevier BV, pp. 248–259. ISSN: 0260-8774.

[391] Kumar Mallikarjuna, Sumanta Pasari and Kamlesh Tiwari. ‘Hierarchical Classification using Neighbourhood Exploration for Sparse Text Tweets’. In: *2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. Vol. 80. IEEE, Jan. 2022, pp. 31–34.

[392] Mangas, Juan J and Moreno, Javier and Picinelli, Anna and Blanco, Domingo. ‘Characterization of cider apple fruits according to their degree of ripening. A chemometric approach’. In: *Journal of Agricultural and Food Chemistry* 46.10 (Oct. 1998). Publisher: American Chemical Society, pp. 4174–4178. ISSN: 0021-8561.

[393] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. *Introduction to Information Retrieval*. USA: Cambridge University Press, July 2008, pp. I–XXI, 1–482. ISBN: 0521865719.

[394] Samuel J. Manoharan. ‘Capsule Network Algorithm for Performance Optimization of Text Classification’. In: *Journal of Soft Computing Paradigm* 3.1 (2021), pp. 1–9.

[395] Yuning Mao et al. ‘Hierarchical Text Classification with Reinforced Label Assignment’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. Ed. by Kentaro Inui et al. Association for Computational Linguistics, Aug. 2019, pp. 445–455.

[396] Matteo Marcuzzo et al. ‘A multi-level approach for hierarchical Ticket Classification’. In: *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*. Gyeongju, Republic of Korea: Association for Computational Linguistics, Oct. 2022, pp. 201–214.

[397] Matteo Marcuzzo et al. ‘Recommendation Systems: An Insight Into Current Development and Future Research Challenges’. In: *IEEE Access* 10 (2022), pp. 86578–86623. ISSN: 2169-3536.

[398] André F. T. Martins and Ramón F. Astudillo. ‘From softmax to sparsemax: a sparse model of attention and multi-label classification’. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*. Ed. by Maria-Florina Balcan and Kilian Q. Weinberger. Vol. abs/1602.02068. Icml’16. New York, NY, USA: JMLR.org, June 2016, pp. 1614–1623. (Visited on 31/05/2021).

[399] Luca Masera and Enrico Blanzieri. ‘AWX: An Integrated Approach to Hierarchical-Multilabel Classification’. In: *MACHINE LEARNING AND KNOWLEDGE DISCOVERY IN DATABASES, ECML PKDD 2018, PT I*. Ed. by M Berlingerio et al. Vol. 11051. Lecture Notes in Artificial Intelligence. European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD), Dublin, IRELAND, SEP 10-14, 2018. Springer International Publishing, 2019, pp. 322–336. ISBN: 978-3-030-10925-7.

[400] Abir Masmoudi et al. ‘A co-training-based approach for the hierarchical multi-label classification of research papers’. In: *Expert Systems* 38.4 (June 2021), e12613. ISSN: 0266-4720.

[401] Soheila Masoudian, Vali Derhami and Sajjad Zarifzadeh. ‘Hierarchical Persian Text Categorization in Absence of Labeled Data’. In: *2019 27th Iranian Conference on Electrical Engineering (ICEE)*. Vol. 3. IEEE, May 2019, pp. 1951–1955.

[402] Matveyeva, Tatiana A. and Sarimov, Ruslan M. and Simakin, Alexander V. and Astashev, Maxim E. and Burmistrov, Dmitriy E. and Lednev, Vasily N. and Sdvizhenskii, Pavel A. and Grishin, Mikhail Ya and Pershin, Sergey M. and Chilingaryan, Narek O. and Semenova, Natalya A. and Dorokhov, Alexey S. and Gudkov, Sergey V. ‘Using Fluorescence Spectroscopy to Detect Rot in Fruit and Vegetable Crops’. In: *Applied Sciences* 12.7 (Jan. 2022). Number: 7 Publisher: Multidisciplinary Digital Publishing Institute, p. 3391. ISSN: 2076-3417. (Visited on 21/02/2023).

[403] Mazen, Fatma M A and Nashat, Ahmed A. ‘Ripeness classification of bananas using an artificial neural network’. In: *Arabian Journal for Science and Engineering* 44.8 (Aug. 2019). Publisher: Springer Science and Business Media LLC, pp. 6901–6910. ISSN: 2191-4281.

[404] Julian McAuley and Jure Leskovec. ‘Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text’. In: *Proceedings of the 7th ACM Conference on Recommender Systems*. Ed. by Qiang Yang et al. RecSys ’13. Hong Kong, China: Association for Computing Machinery, Oct. 2013, pp. 165–172. ISBN: 9781450324090.

[405] McGlone, V Andrew and Fraser, Daniel G and Jordan, Robert B and Künnemeyer, Rainer. ‘Internal quality assessment of mandarin fruit by vis/NIR spectroscopy’. In: *Journal of Near Infrared Spectroscopy* 11.5 (Oct. 2003). Publisher: SAGE Publications, pp. 323–332. ISSN: 0967-0335.

[406] Scott Mayer McKinney et al. ‘International evaluation of an AI system for breast cancer screening’. In: *Nature* 577.7788 (Jan. 2020), pp. 89–94. ISSN: 0028-0836.

[407] Kyrylo Medianovskyi and Ahti-Veikko Pietarinen. ‘On Explainable AI and Abductive Inference’. en. In: *Philosophies* 7.2 (Mar. 2022), p. 35. ISSN: 2409-9287.

[408] Medlicott, A P and Reynolds, S B and Thompson, A K. ‘Effects of temperature on the ripening of mango fruit (mangifera indica l, var. tommy atkins)’. In: *Journal of the Science of Food and Agriculture* 37.5 (May 1986), pp. 469–474. ISSN: 0022-5142.

[409] MedlinePlus. *Neuromuscular Disorders*. `https://medlineplus.gov/neuromusculardisorders.html`. Sept. 2008.

[410] Muhammad Mehar et al. ‘Understanding a Revolutionary and Flawed Grand Experiment in Blockchain: The DAO Attack’. In: *Journal of Cases on Information Technology* 21.1 (2019), pp. 19–32.

[411] Clara Meister et al. ‘Is Sparse Attention more Interpretable?’ In: *CoRR* abs/2106.01087 (June 2021). Ed. by Chengqing Zong et al., pp. 122–129. arXiv: 2106.01087.

[412] Giovanna Menardi and Nicola Torelli. ‘Training and assessing classification rules with imbalanced data’. In: *Data mining and knowledge discovery* 28 (Jan. 2014), pp. 92–122. ISSN: 1384-5810.

[413] Corrado Mencar. ‘Interpretability of Fuzzy Systems’. In: *Fuzzy Logic and Applications*. Ed. by Francesco Masulli, Gabriella Pasi and Ronald R. Yager. Vol. 8256. Springer International Publishing, Nov. 2013, pp. 22–35. ISBN: 978-3319031996.

[414] Eneldo Loza Mencía and Johannes Fürnkranz. ‘Efficient Pairwise Multilabel Classification for Large-Scale Problems in the Legal Domain’. In: *Proceedings of the 2008th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part II*. Ed. by Walter Daelemans, Bart Goethals and Katharina Morik. Vol. 5212. ECMLPKDD’08. Antwerp, Belgium: Springer-Verlag, Aug. 2008, pp. 50–65. ISBN: 3540874801.

[415] Mendoza, F and Lu, R and Ariana, D and Cen, H and Bailey, B. ‘Integrated spectral and image analysis of hyperspectral scattering data for prediction of apple fruit firmness and soluble solids content’. In: *Postharvest Biology and Technology* 62 (July 2011), pp. 149–160. ISSN: 0925-5214.

[416] Mendoza, F. and Aguilera, J. M. ‘Application of Image Analysis for Classification of Ripening Bananas’. In: *Journal of Food Science* 69.9 (Dec. 2004). _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1365-2621.2004.tb09932.x, E471–E477. ISSN: 1750-3841. (Visited on 21/02/2023).

[417] Kevin Meng et al. ‘Locating and editing factual associations in GPT’. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Ed. by Sanmi Koyejo et al. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., Feb. 2022. ISBN: 9781713871088.

[418] Lili Meng et al. ‘Interpretable Spatio-Temporal Attention for Video Action Recognition’. In: *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. Issn: 2473-9944. IEEE, Oct. 2019, pp. 1513–1522.

[419] Yu Meng et al. ‘Weakly-Supervised Hierarchical Text Classification’. In: *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*. Vol. 33. 01. AAAI Press, July 2019, pp. 6826–6833.

[420] Mercado-Silva, Edmundo and Benito-Bautista, Pedro and de los Angeles García-Velasco, Ma. ‘Fruit development, harvest index and ripening changes of guavas produced in central Mexico’. In: *Postharvest Biology and Technology* 13.2 (May 1998), pp. 143–150. ISSN: 0925-5214. (Visited on 21/02/2023).

[421] Stefano Mezza, Wayne Wobcke and Alan Blair. ‘A Multi-Dimensional, Cross-Domain and Hierarchy-Aware Neural Architecture for ISO-Standard Dialogue Act Tagging’. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Ed. by Nicoletta Calzolari et al. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, Oct. 2022, pp. 542–552.

[422] Sabrina J. Mielke et al. ‘Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP’. In: *arXiv* abs/2112.10508 (Dec. 2021). arXiv (preprint): 2112.10508 (cs.CL).

[423] Tomas Mikolov et al. ‘Distributed Representations of Words and Phrases and their Compositionality’. In: *Advances in Neural Information Processing Systems 26*. Ed. by Christopher J. C. Burges et al. Vol. 26. NIPS’13. Curran Associates, Inc., Oct. 2013, pp. 3111–3119.

[424] Tomas Mikolov et al. ‘Efficient Estimation of Word Representations in Vector Space’. In: *1st International Conference on Learning Representations, ICLR 2013*. Vol. abs/1301.3781. Cornell University, May 2013. arXiv: 1301.3781 [cs.CL].

[425] Tim Miller. ‘Explanation in artificial intelligence: Insights from the social sciences’. In: *Artificial intelligence* 267 (Feb. 2019), pp. 1–38. ISSN: 0004-3702.

[426] Shervin Minaee et al. ‘Deep Learning–Based Text Classification: A Comprehensive Review’. In: *Acm Comput. Surv.* 54.3 (May 2021), pp. 1–40. ISSN: 0360-0300.

[427] Miraei Ashtiani, Seyed-Hassan and Javanmardi, Shima and Jahanbanifard, Mehrdad and Martynenko, Alex and Verbeek, Fons J. ‘Detection of Mulberry Ripeness Stages Using Deep Learning Models’. In: *IEEE Access* 9 (2021). Conference Name: IEEE Access, pp. 100380–100394. ISSN: 2169-3536.

[428] Brent Mittelstadt, Chris Russell and Sandra Wachter. ‘Explaining Explanations in AI’. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*. FAT*’19. Atlanta, GA, USA: Association for Computing Machinery, Jan. 2019, pp. 279–288.

[429] Mollazade, Kaveh and Omid, Mahmoud and Tab, Fardin Akhlaghian and Mohtasebi, Sayed Saeid. ‘Principles and applications of light backscattering imaging in quality evaluation of Agro-food products: A review’. In: *Food and Bioprocess Technology* 5.5 (July 2012). Publisher: Springer Science and Business Media LLC, pp. 1465–1485. ISSN: 1935-5130.

[430] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. 2nd ed. Independently published (February 28, 2022), 2022.

[431] Shane T Mueller et al. ‘Explanation in Human-AI Systems: A Literature Meta-Review, Synopsis of Key Ideas and Publications, and Bibliography for Explainable AI’. In: *arXiv (Cornell University)* abs/1902.01876 (Feb. 2019). ISSN: 2331-8422. arXiv: 1902.01876 [cs.AI].

[432] Muhua, L and Peng, F and Renfa, C. ‘Non-destructive estimation peach ssc and firmness by mutispectral reflectance imaging’. In: *New Zealand Journal of Agricultural Research* 50 (Dec. 2007), pp. 601–608. ISSN: 0028-8233.

[433] James Mullenbach et al. ‘Explainable Prediction of Medical Codes from Clinical Text’. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji and Amanda Stent. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 1101–1111. (Visited on 31/05/2021).

[434] Silvia Multari et al. ‘Predicting Metabolic Reactions with a Molecular Transformer for Drug Design Optimization’. In: *2024 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE. IEEE, Aug. 2024, pp. 1–8.

[435] Munawar, Agus Arip and Kusumiyati and Wahyuni, Devi. ‘Near infrared spectroscopic data for rapid and simultaneous prediction of quality attributes in intact mango fruits’. In: *Data in Brief* 27 (Dec. 2019), p. 104789. ISSN: 2352-3409.

[436] Munoz, C and Avila, J and Salvo, S and Huircan, J I. ‘Prediction of harvest start date in highbush blueberry using time series regression models with correlated errors’. In: *Scientia Horticulturae* 138 (May 2012), pp. 165–170. ISSN: 0304-4238.

[437] W James Murdoch et al. ‘Definitions, methods, and applications in interpretable machine learning’. en. In: *Proc. Natl. Acad. Sci. U. S. A.* 116.44 (Oct. 2019), pp. 22071–22080. ISSN: 0027-8424.

[438] Nagata, M. and Tallada, J. and Ishino, F. and Gejima, Y. and Kai, S. ‘Estimation of Tomato Ripening Stages Using Three Color Models’. In: *Bulletin of the Faculty of Agriculture, Miyazaki University* 50.1 (Mar. 2004), pp. 65–72. ISSN: 0544-6066. (Visited on 21/02/2023).

[439] Mohammad Naiseh et al. ‘Explainable recommendation: when design meets trust calibration’. In: *World Wide Web* 24.5 (2021), pp. 1857–1884. ISSN: 1573-1413. DOI: 10.1007/s11280-021-00916-0. URL: https://doi.org/10.1007/s11280-021-00916-0.

[440] Felipe Kenji Nakano, Ricardo Cerri and Celine Vens. ‘Active learning for hierarchical multi-label classification’. In: *Data Mining and Knowledge Discovery* 34.5 (Sept. 2020), pp. 1496–1530. ISSN: 1573-756X.

[441] Nambi, E and Kulandasamy, T and Jesudas, M. *Scientific classification of ripening period and development of colourgrade chart for indian mangoes (mangifera indica l.) using multivariate cluster analysis*. 2015.

[442] Author Names. ‘A Theoretical Framework for AI Models Explainability with Applications in Biomedicine’. In: *ArXiv* (2022).

[443] Author Names. ‘An Introduction to Machine Learning Approaches for Biomedicine’. In: *Front. Med.* 8 (2021), p. 771607.

[444] Author Names. ‘Explainable Artificial Intelligence (XAI) in Biomedicine’. In: *MDPI* 2.1 (2023).

[445] Author Names. ‘How Machine Learning will Transform Biomedicine’. In: *PMC* (2020).

[446] Author Names. ‘Machine Learning Approaches to Retrieve High-Quality Clinical Articles in the Biomedical Literature’. In: *PMC* (2021).

[447] Neel Nanda et al. *Progress measures for grokking via mechanistic interpretability*. Jan. 2023.

[448] Manish Narwaria. ‘Does explainable machine learning uncover the black box in vision applications?’ In: *Image and Vision Computing* 118 (Feb. 2022), p. 104353. ISSN: 0262-8856.

[449] Nadia Nashid, Thareq Sistematiko and Shajedul Alam Chowdhury. ‘Can large language models write good code? a large-scale empirical study’. In: *arXiv preprint arXiv:2311.18249* (2023). arXiv: 2311.18249 [cs.SE].

[450] Meike Nauta et al. ‘This Looks Like That, Because ... Explaining Prototypes for Interpretable Image Recognition’. In: *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*. Ed. by Michael Kamp et al. Vol. 1524. Springer International Publishing, 2021, pp. 441–456. ISBN: 978-3030937355.

[451] Neimark, Daniel and Bar, Omri and Zohar, Maya and Asselmann, Dotan. ‘Video Transformer Network’. In: *arXiv preprint arXiv:2102.00719* (Oct. 2021), pp. 3156–3165.

[452] Anh Nguyen, Jason Yosinski and Jeff Clune. ‘Understanding Neural Networks via Feature Visualization: A Survey’. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Ed. by Wojciech Samek et al. Vol. 11700. Cham: Springer International Publishing, May 2019, pp. 55–76. ISBN: 978-3-030-28954-6.

[453] Thanh-Tung Nguyen et al. ‘A Two-Stage Decoder for Efficient ICD Coding’. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 4658–4665.

[454] Jianmo Ni, Jiacheng Li and Julian McAuley. ‘Justifying Recommendations using Distantly-Labeled Reviews and Fine-Grained Aspects’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 188–197.

[455] Ni, Xueping and Li, Changying and Jiang, Huanyu and Takeda, Fumiomi. ‘Deep learning image segmentation and extraction of blueberry fruit traits associated with harvestability and yield’. In: *Horticulture Research* 7.1 (July 2020), p. 110. ISSN: 2052-7276.

[456] Nicolai Häni and Pravakar Roy and Volkan Isler. ‘MinneApple: A Benchmark Dataset for Apple Detection and Segmentation’. In: *CoRR* abs/1909.06441 (Sept. 2019), pp. 852–858. ISSN: 2377-3766. arXiv: 1909.06441.

[457] Ian E. Nielsen et al. ‘Robust Explainability: A tutorial on gradient-based attribution methods for deep neural networks’. In: *IEEE Signal Processing Magazine* 39.4 (July 2022), pp. 73–84. ISSN: 1053-5888.

[458] Ivica Nikolić et al. ‘Finding the greedy, prodigal, and suicidal contracts at scale’. In: *Proceedings of the 34th Annual Computer Security Applications Conference*. ACM, Dec. 2018, pp. 653–663.

[459] Bo Ning et al. ‘UMP-MG: A Uni-directed Message-Passing Multi-label Generation Model for Hierarchical Text Classification’. In: *Data Science and Engineering* 8.2 (June 2023), pp. 112–123. ISSN: 2364-1541.

[460] Marco S Nobile et al. ‘Unsupervised neural networks as a support tool for pathology diagnosis in MALDI-MSI experiments: A case study on thyroid biopsies’. In: *Expert Systems with Applications* 215 (May 2023), p. 119296. ISSN: 0957-4174.

[461] Harsha Nori et al. *InterpretML: A Unified Framework for Machine Learning Interpretability*. Sept. 2019.

[462] Helen O’Brien Quinn et al. ‘Literature review of explainable tabular data analysis’. In: *Electronics* 13.19 (Sept. 2024), p. 3806. ISSN: 2079-9292.

[463] Chris Olah et al. ‘Zoom In: An Introduction to Circuits’. In: *Distill* 5.3 (Mar. 2020). ISSN: 2476-0757.

[464] Olarewaju, O O and Bertling, I and Magwaza, L S. ‘Non-destructive evaluation of avocado fruit maturity using near infrared spectroscopy and PLS regression models’. In: *Scientia Horticulturae* 199 (Feb. 2016). Publisher: Elsevier BV, pp. 229–236. ISSN: 0304-4238.

[465] Olmo, M and Nadas, A and Garcia, J M. ‘Nondestructive methods to evaluate maturity level of oranges’. In: *Journal of Food Science* 65.2 (Mar. 2000). Publisher: Wiley, pp. 365–369. ISSN: 0022-1147.

[466] Opara, Linus U and Al-Said, Fahad A and Al-Abri, Aamna. ‘Assessment of what the consumer values in fresh fruit quality: Case study of Oman’. In: *New Zealand Journal of Crop and Horticultural Science* 35.2 (June 2007). Publisher: Informa UK Limited, pp. 235–243. ISSN: 0114-0671.

[467] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[468] Oxford Advanced American Dictionary. *interpret verb - Definition, pictures, pronunciation and usage notes*. en. https://www.oxfordlearnersdictionaries.com/definition/american_english/interpret. Accessed: 2024-2-16. 2024.

[469] Ozkan, Ilker Ali and Koklu, Murat and Saraçoğlu, Rıdvan. ‘Classification of Pistachio Species Using Improved K-NN Classifier’. In: *Progress in Nutrition* 23.2 (July 2021), e2021044.

[470] Ozkan, Ilker Ali and Koklu, Murat and Saraçoğlu, Rıdvan. *Pistachio Image Dataset*. 2022.

[471] Ankit Pal., Muru Selvakumar. and Malaikannan Sankarasubbu. ‘MAGNET: Multi-Label Text Classification using Attention-based Graph Neural Network’. In: *Proceedings of the 12th International Conference on Agents and Artificial Intelligence - Volume 2: ICAART,* ed. by Ana Paula Rocha, Luc Steels and H. Jaap van den Herik. Vol. abs/2003.11644. INSTICC. SciTePress, Mar. 2020, pp. 494–505. ISBN: 978-989-758-395-7.

[472] Cecilia Panigutti et al. ‘Co-design of Human-centered, Explainable AI for Clinical Decision Support’. In: *ACM Trans. Interact. Intell. Syst.* 13.4 (Dec. 2023). ISSN: 2160-6455. DOI: 10.1145/3587271. URL: https://doi.org/10.1145/3587271.

[473] Authors of The Mdpi 2023 Paper. ‘Exploring Evaluation Methods for Interpretable Machine Learning’. In: *MDPI* 14.8 (2023).

[474] Daniele M Papetti et al. ‘An accurate and time-efficient deep learning-based system for automated segmentation and reporting of cardiac magnetic resonance-detected ischemic scar’. In: *Computer methods and programs in biomedicine* 229 (Feb. 2023), p. 107321. ISSN: 0169-2607.

[475] Pardede, Jasman and Husada, Milda Gustiana and Hermana, Asep Nana and Rumapea, Sri Agustina. ‘Fruit Ripeness Based on RGB, HSV, HSL, L*a*b*Color Feature Using SVM’. In: *2019 International Conference of Computer Science and Information Technology (ICoSNIKOM)*. IEEE, Nov. 2019, pp. 1–5.

[476] Jihyun Park, Sunkyu Lee and Minjoon Kim. ‘Hybrid Reasoning: Fusing Retrieval and Step-by-Step Thinking in Large Language Models’. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*. 2024, pp. 4501–4515.

[477] Razvan Pascanu, Tomas Mikolov and Yoshua Bengio. ‘On the difficulty of training recurrent neural networks’. In: *Proceedings of the 30th International Conference on Machine Learning*. Ed. by Sanjoy Dasgupta and David McAllester. Vol. 28. Proceedings of Machine Learning Research. Atlanta, Georgia, USA: PMLR, June 2013, pp. 1310–1318.

[478] Vivek Patel et al. ‘Patient Engagement in the Digital Age’. In: *Patient Educ. Couns.* (2019).

[479] Judea Pearl and Dana Mackenzie. *The Book of Why: The New Science of Cause and Effect*. 1st. USA: Basic Books, Inc., May 2018. ISBN: 046509760X.

[480] Charles Sanders Peirce. *Collected Papers of Charles Sanders Peirce*. en. Harvard University Press, 1974.

[481] Peirs, A and Lammertyn, J and Ooms, K and Nicolaı, B M. ‘Prediction of the optimal picking date of different apple cultivars by means of VIS/NIR-spectroscopy’. In: *Postharvest Biology and Technology* 21.2 (Jan. 2001), pp. 189–199. ISSN: 0925-5214.

[482] Peirs, A and Scheerlinck, N and Nicolaı, B M. ‘Temperature compensation for near infrared reflectance measurement of apple fruit soluble solids contents’. In: *Postharvest Biology and Technology* 30.3 (Dec. 2003), pp. 233–248. ISSN: 0925-5214.

[483] Hao Peng et al. ‘Hierarchical Taxonomy-Aware and Attentional Graph Capsule RCNNs for Large-Scale Multi-Label Text Classification’. In: *IEEE Trans. Knowl. Data Eng.* 33.6 (June 2021), pp. 2505–2519. ISSN: 1041-4347.

[484] Hao Peng et al. ‘Large-Scale Hierarchical Text Classification with Recursively Regularized Deep Graph-CNN’. In: *Proceedings of the 2018 World Wide Web Conference*. Ed. by Pierre-Antoine Champin et al. WWW ’18. Lyon, France: International World Wide Web Conferences Steering Committee, May 2018, pp. 1063–1072. ISBN: 9781450356398.

[485] Peng, Yankun and Lu, Renfu. ‘Analysis of spatially resolved hyperspectral scattering images for assessing apple fruit firmness and soluble solids content’. In: *Postharvest Biology and Technology* 48.1 (May 2008). Publisher: Elsevier BV, pp. 52–62. ISSN: 0925-5214.

[486] Jeffrey Pennington, Richard Socher and Christopher Manning. ‘GloVe: Global Vectors for Word Representation’. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Ed. by Alessandro Moschitti, Bo Pang and Walter Daelemans. Association for Computational Linguistics, Oct. 2014, pp. 1532–1543.

[487] Matthew E. Peters et al. ‘Deep Contextualized Word Representations’. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Ed. by Marilyn A. Walker, Heng Ji and Amanda Stent. Vol. abs/1802.05365. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237.

[488] Hieu Pham et al. ‘A multi-head attention-like feature selection approach for tabular data’. In: *Knowledge-Based Systems* 301 (Oct. 2024), p. 112250. ISSN: 0950-7051.

[489] Martha C Piñeros-Fernández. ‘Artificial intelligence applications in the diagnosis of neuromuscular diseases: a narrative review’. In: *Cureus* 15.11 (Nov. 2023). ISSN: 2168-8184.

[490] Mara Pistellato et al. ‘Adaptive Albedo Compensation for Accurate Phase-Shift Coding’. In: *2018 24th International Conference on Pattern Recognition (ICPR)*. Vol. 2. IEEE Computer Society, Aug. 2018, pp. 2450–2455.

[491] Stephen M. Pizer et al. ‘Adaptive histogram equalization and its variations’. In: *Computer Vision, Graphics, and Image Processing* 39.3 (Sept. 1987), pp. 355–368. ISSN: 0734-189X.

[492] S. Popov, S. Morozov and A. Babenko. ‘Neural oblivious decision ensembles for deep learning on tabular data’. In: *Advances in Neural Information Processing Systems*. Vol. 32. Cornell University, Sept. 2019.

[493] Faizal Adhitama Prabowo, Muhammad Okky Ibrohim and Indra Budi. ‘Hierarchical Multi-label Classification to Identify Hate Speech and Abusive Language on Indonesian Twitter’. In: *2019 6th International Conference on Information Technology, Computer and Electrical Engineering (ICITACEE)*. IEEE, Sept. 2019, pp. 1–5.

[494] W Nicholson Price 2nd and I Glenn Cohen. ‘Privacy in the age of medical big data’. en. In: *Nat. Med.* 25.1 (Jan. 2019), pp. 37–43. ISSN: 1078-8956.

[495] V. Prinet, D. Lischinski and M. Werman. ‘Illuminant Chromaticity from Image Sequences’. In: *2013 IEEE International Conference on Computer Vision*. IEEE, Dec. 2013, pp. 3320–3327.

[496] Subhash Chandra Pujari, Annemarie Friedrich and Jannik Strötgen. ‘A Multi-task Approach to Neural Multi-label Hierarchical Patent Classification Using Transformers’. In: *Advances in Information Retrieval*. Ed. by Djoerd Hiemstra et al. Vol. 12656. Cham: Springer International Publishing, 2021, pp. 513–528. ISBN: 978-3-030-72113-8.

[497] Kunal Punera and Joydeep Ghosh. ‘Enhanced Hierarchical Classification via Isotonic Smoothing’. In: *Proceedings of the 17th International Conference on World Wide Web*. Ed. by Jinpeng Huai et al. WWW ’08. Beijing, China: Association for Computing Machinery, May 2008, pp. 151–160. ISBN: 9781605580852.

[498] Peng Qian et al. ‘Cross-modality mutual learning for enhancing smart contract vulnerability detection on bytecode’. In: *Proceedings of the ACM Web Conference 2023*. Ed. by Ying Ding et al. ACM, May 2023, pp. 2220–2229.

[499] Y. Qian et al. ‘Flash Lightens Gray Pixel’. In: *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, Aug. 2019, pp. 4604–4608.

[500] Y. Qian et al. ‘Recurrent Color Constancy’. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, Oct. 2017, pp. 5459–5467.

[501] Yanlin Qian et al. *A Benchmark for Temporal Color Constancy*. arXiv: 2003.03763. Mar. 2020. arXiv: `2003.03763 [cs.CV]`. (Visited on 31/05/2021).

[502] Qin, J and Lu, R and Peng, Y. ‘Prediction of apple internal quality using spectral absorption and scattering properties’. In: *Transactions of the ASABE* 52.2 (2009), pp. 489–496. ISSN: 2151-0032.

[503] J. Qiu, H. Xu and Z. Ye. ‘Color Constancy by Reweighting Image Feature Maps’. In: *IEEE Transactions on Image Processing* 29 (2020), pp. 5711–5721. ISSN: 1057-7149.

[504] Qiu, Q. and Shi, K. and Qiao, X. J. and Jiang, K. ‘Determining the Dominant Environmental Parameters for Greenhouse Tomato Seedling Growth Modeling Using Canonical Correlation Analysis’. In: *IFAC-PapersOnLine*. 5th IFAC Conference on Sensing, Control and Automation Technologies for Agriculture AGRICONTROL 2016 49.16 (Jan. 2016), pp. 387–391. ISSN: 2405-8963. (Visited on 21/02/2023).

[505] Yuhui Quan et al. ‘Attention with structure regularization for action recognition’. en. In: *Computer Vision and Image Understanding* 187 (Oct. 2019), p. 102794. ISSN: 1077-3142. (Visited on 31/05/2021).

[506] Alec Radford et al. *Improving Language Understanding by Generative Pre-Training*. 2018.

[507] Alec Radford et al. *Language Models are Unsupervised Multitask Learners*. 2019.

[508] Radiopaedia. *MRI sequences (overview)*. `https://radiopaedia.org/articles/mri-sequences-overview`.

[509] Ragni, Luigi and Cevoli, Chiara and Berardinelli, Annachiara and Silaghi, Florina Aurelia. ‘Non-destructive internal quality assessment of “Hayward” kiwifruit by waveguide spectroscopy’. In: *Journal of Food Engineering* 109.1 (Mar. 2012), pp. 32–37. ISSN: 0260-8774.

[510] A. Rahimi and B. Recht. ‘Random features for large-scale kernel machines’. In: *Advances in Neural Information Processing Systems*. Ed. by John C. Platt et al. Vol. 20. MIT Press, Dec. 2007, pp. 1177–1184.

[511] Rajkumar, P and Wang, N and EImasry, G and Raghavan, G S V and Gariepy, Y. ‘Studies on banana fruit quality and maturity stages using hyperspectral imaging’. In: *Journal of Food Engineering* 108.1 (Jan. 2012). Publisher: Elsevier BV, pp. 194–200. ISSN: 0260-8774.

[512] Rajeev Ramanath et al. ‘Color image processing pipeline’. In: *Signal Processing Magazine, IEEE* 22.1 (Feb. 2005). Conference Name: IEEE Signal Processing Magazine, pp. 34–43. ISSN: 1558-0792.

[513] Randhawa, H S and Sharma, S and Student, C S E. ‘A survey of computer vision and soft computing techniques for ripeness grading of fruits’. In: *Journal of Advanced Computing and Communication Technologies* 2 (2014).

[514] Nils Reimers and Iryna Gurevych. ‘Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. Vol. abs/1908.10084. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 3982–3992.

[515] Weijieying Ren et al. ‘Deep Learning within Tabular Data: Foundations, Challenges, Advances and Future Directions’. In: *arXiv preprint arXiv:2501.03540* abs/2501.03540 (Jan. 2025). ISSN: 2331-8422.

[516] Ren, Shaoqing and He, Kaiming and Girshick, Ross and Sun, Jian. ‘Towards Real-Time Object Detection with Region Proposal Networks’. In: *Advances in Neural Information Processing Systems 28*. Ed. by C. Cortes et al. Curran Associates, Inc., 2015, pp. 942–950.

[517] D. Ressi et al. ‘Reentrancy detection tools in the age of LLMs’. In: *ACM International Conference on the Foundations of Software Engineering (FSE ’26)*. Manuscript submitted for publication. 2026.

[518] D. Ressi et al. ‘SoK: Benchmarking Failure — The State (and Decay) of Reentrancy Detection Tools’. In: *47th IEEE Symposium on Security and Privacy (S&P ’26)*. Manuscript submitted for publication. 2026.

[519] Dalila Ressi et al. ‘Vulnerability Detection in Ethereum Smart Contracts via Machine Learning: A Qualitative Analysis’. In: *arXiv preprint arXiv:2407.18639* abs/2407.18639 (July 2024). ISSN: 2331-8422. arXiv: 2407.18639 [cs.CR].

[520] Marco Tulio Ribeiro, Sameer Singh and Carlos 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*. Ed. by John DeNero, Mark Finlayson and Sravana Reddy. KDD ’16. San Francisco, California, USA: Assoc. for Computing Machinery, Feb. 2016, pp. 1135–1144. ISBN: 9781450342322.

[521] C. Rinaldi et al. ‘The genesis of twin transition: Understanding the evolution of the digital and sustainable transitions’. In: *International Journal of Information Management* (2025). Manuscript submitted for publication.

[522] Rinnan, Asmund and van den Berg, Frans and Engelsen, Søren Balling. ‘Review of the most common pre-processing techniques for near-infrared spectra’. In: *TrAC Trends in Analytical Chemistry* 28.10 (Nov. 2009). Publisher: Elsevier BV, pp. 1201–1222. ISSN: 0165-9936.

[523] Julian Risch, Samuele Garda and Ralf Krestel. ‘Hierarchical Document Classification as a Sequence Generation Task’. In: *Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020*. Ed. by Ruhua Huang et al. JCDL ’20. Virtual Event, China: Association for Computing Machinery, Aug. 2020, pp. 147–155. ISBN: 9781450375856.

[524] Rivero Mesa, Armacheska and Chiang, John. ‘Non-invasive Grading System for Banana Tiers using RGB Imaging and Deep Learning’. In: *2021 7th International Conference on Computing and Artificial Intelligence*. ICCAI 2021. event-place: Tianjin, China. New York, NY, USA: Association for Computing Machinery, Sept. 2021, pp. 113–118.

[525] T. J. Rivlin. *The Chebyshev Polynomials*. Wiley-Interscience, 1974.

[526] M. Rizzo et al. ‘A comparison of machine learning techniques for Ethereum smart contract vulnerability detection’. In: *CEUR Workshop Proceedings*. Ed. by D. Porello, C. Vinci and M. Zavatteri. Vol. 3904. CEUR-WS.org, 2024, pp. 119–126. URL: https://ceur-ws.org/Vol-3904/paper15.pdf.

[527] M. Rizzo et al. ‘A theoretical framework for AI models explainability with application in biomedicine’. In: *2023 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, Aug. 2023, pp. 1–9. DOI: 10.1109/CIBCB56990.2023.10264877. URL: https://arxiv.org/pdf/2212.14447.

[528] M. Rizzo et al. ‘Evaluating the faithfulness of causality in saliency-based explanations of deep learning models for temporal colour constancy’. In: *Explainable Artificial Intelligence*. Ed. by L. Longo, S. Lapuschkin and C. Seifert. Vol. 2155. Cham: Springer, 2024, pp. 125–142. ISBN: 978-3031637995. DOI: 10.1007/978-3-031-63800-8_7. URL: https://doi.org/10.1007/978-3-031-63800-8_7.

[529] M. Rizzo et al. ‘Fruit ripeness classification: A survey’. In: *Artificial Intelligence in Agriculture* 7 (Mar. 2023), pp. 44–57. ISSN: 2589-7217. DOI: 10.1016/j.aiia.2023.02.004. URL: https://doi.org/10.1016/j.aiia.2023.02.004.

[530] M. Rizzo et al. ‘Leveraging periodicity for tabular deep learning’. In: *Electronics* 14.6 (Mar. 2025), p. 1165. ISSN: 2079-9292. DOI: 10.3390/electronics14061165. URL: https://doi.org/10.3390/electronics14061165.

[531] M. Rizzo et al. ‘Stop overkilling simple tasks with black-box models, use more transparent models instead’. In: *Pattern Recognition and Artificial Intelligence*. Ed. by C. Wallraven, C.-L. Liu and A. Ross. Singapore: Springer, 2025, pp. 279–293. ISBN: 978-9819787012. DOI: 10.1007/978-981-97-8702-9_19. URL: https://doi.org/10.1007/978-981-97-8702-9_19.

[532] Matteo Rizzo et al. ‘Advanced Large Language Models Prompting Strategies for Reentrancy Classification and Explanation in Smart Contracts’. In: *Blockchain Technology and Emerging Applications*. Ed. by William Knottenbelt et al. Cham: Springer Nature Switzerland, 2026, pp. 37–56. ISBN: 978-3-032-12335-0.

[533] Matteo Rizzo et al. ‘Cascading Convolutional Temporal Colour Constancy’. In: *CoRR* abs/2106.07955 (June 2021), p. 013049. arXiv: 2106.07955.

[534] Matteo Rizzo et al. ‘Machine learning models explanations as interpretations of evidence: a theoretical framework of explainability and its implications on high-stakes biomedical decision-making’. In: *BMC Medical Research Methodology* 25.Suppl 1 (2025), p. 282. DOI: 10.1186/s12874-025-02703-1.

[535] Kervy Rivas Rojas et al. ‘Efficient Strategies for Hierarchical Text Classification: External Knowledge and Auxiliary Tasks’. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Vol. abs/2005.02473. Association for Computational Linguistics, May 2020, pp. 2252–2257.

[536] Miguel Romero, Jorge Finke and Camilo Rocha. ‘A top-down supervised learning approach to hierarchical multi-label classification in networks’. In: *Applied Network Science* 7.1 (Feb. 2022), p. 8. ISSN: 2364-8228.

[537] Andrew Ross and Finale Doshi-Velez. ‘Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing Their Input Gradients’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 32.1 (Apr. 2018). Ed. by Sheila A. McIlraith and Kilian Q. Weinberger. ISSN: 2159-5399.

[538] Georgios Rovolis and Abdolrasoul Habibipour. ‘When participatory design meets data-driven decision making: A literature review and the way forward’. In: *Manag. Sci. Lett.* 14.2 (2024), pp. 107–126. ISSN: 1923-9335.

[539] Royer, C A. ‘Fluorescence spectroscopy’. In: *Methods in Molecular Biology* 40 (1995), pp. 65–89.

[540] Cynthia Rudin. ‘Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead’. en. In: *Nat Mach Intell* 1.5 (May 2019), pp. 206–215. ISSN: 2522-5839.

[541] 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*. Cambridge, MA, USA: MIT Press, Jan. 1986. Chap. 11, pp. 318–362. ISBN: 026268053X.

[542] Sa, Inkyu and Ge, Zongyuan and Dayoub, Feras and Upcroft, Ben and Perez, Tristan and McCool, Chris. ‘DeepFruits: A Fruit Detection System Using Deep Neural Networks’. en. In: *Sensors* 16.8 (Aug. 2016). Number: 8 Publisher: Multidisciplinary Digital Publishing Institute, p. 1222. ISSN: 1424-8220. (Visited on 21/02/2023).

[543] Sara Sabour, Nicholas Frosst and Geoffrey E. Hinton. ‘Dynamic Routing between Capsules’. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Ed. by Isabelle Guyon et al. Vol. abs/1710.09829. NIPS’17. Long Beach, California, USA: Curran Associates Inc., Dec. 2017, pp. 3859–3869. ISBN: 9781510860964.

[544] Mobashir Sadat and Cornelia Caragea. ‘Hierarchical Multi-Label Classification of Scientific Documents’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva and Yue Zhang. Vol. abs/2211.02810. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 8923–8937.

[545] S.R. Safavian and D. Landgrebe. ‘A survey of decision tree classifier methodology’. In: *IEEE Transactions on Systems, Man, and Cybernetics* 21.3 (June 1991), pp. 660–674. ISSN: 0018-9472.

[546] Sourav Saha, Debapriyo Majumdar and Mandar Mitra. *Explainability of Text Processing and Retrieval Methods: A Critical Survey*. arXiv:2212.07126 [cs]. Dec. 2022. (Visited on 18/06/2024).

[547] Maria Sahakyan, Zeyar Aung and Talal Rahwan. ‘Explainable Artificial Intelligence for Tabular Data: A Survey’. In: *IEEE Access* 9 (2021), pp. 135392–135422. ISSN: 2169-3536.

[548] Vivien Sainte Fare Garnot and Loic Landrieu. ‘Leveraging Class Hierarchies with Metric-Guided Prototype Learning’. In: *32th British Machine Vision Conference*. Online: BMVA Press, 2021, p. 123.

[549] Elizabeth Salesky, David Etter and Matt Post. ‘Robust Open-Vocabulary Translation from Visual Text Representations’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Vol. abs/2104.08211. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, May 2021, pp. 7235–7252.

[550] Wojciech Samek. ‘Chapter 2 - Explainable deep learning: concepts, methods, and new developments’. In: *Explainable Deep Learning AI*. Ed. by Jenny Benois-Pineau et al. Academic Press, 2023, pp. 7–33. ISBN: 978-0-323-96098-4.

[551] Dvir Samuel, Yuval Atzmon and Gal Chechik. ‘From generalized zero-shot learning to long-tail with class descriptors’. In: *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, Jan. 2021, pp. 286–295.

[552] Sanchez, M-T and Haba, M J and Benıtez-Lopez, M and Fernandez-Novales, J and Garrido-Varo, A and Perez-Marın, D. ‘Non-destructive characterization and quality control of intact strawberries based on nir spectral data’. In: *Journal of Food Engineering* 110 (May 2012), pp. 102–108. ISSN: 0260-8774.

[553] Evan Sandhaus. *The New York Times Annotated Corpus LDC2008T19*. https://doi.org/10.35111/77ba-9x74. Philadelphia: Linguistic Data Consortium. 2008.

[554] Mark Sandler et al. ‘MobileNetV2: Inverted Residuals and Linear Bottlenecks’. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE Computer Society, June 2018, pp. 4510–4520.

[555] A Sans. ‘Abductive reasoning on humans/AI interactions in medical contexts’. In: *European Journal of Public Health*. Vol. 30. Supplement_5. Oxford University Press (OUP), Sept. 2020.

[556] Saranwong, Innapa and Sornsrivichai, Jinda and Kawano, Sumio. ‘On-tree evaluation of harvesting quality of mango fruit using a hand-held NIR instrument’. In: *Journal of Near Infrared Spectroscopy* 11.4 (Aug. 2003). Publisher: SAGE Publications, pp. 283–293. ISSN: 0967-0335.

[557] Saranya, N and Srinivasan, K and Kumar, S K Pravin. ‘Banana ripeness stage identification: a deep learning approach’. In: *Journal of Ambient Intelligence and Humanized Computing* 13.8 (Aug. 2022). Publisher: Springer Science and Business Media LLC, pp. 4033–4039. ISSN: 1868-5145.

[558] Satpute, M R and Jagdale, S M. ‘Color, size, volume, shape and texture feature extraction techniques for fruits: a review’. In: *International Research Journal of Engineering and Technology* 3 (2016), pp. 703–708.

[559] Satpute, Manali R. and Jagdale, Sumati M. ‘Color, Size, Volume, Shape and Texture Feature Extraction Techniques for Fruits: A Review’. In: 03.04 ().

[560] Marco Savarese et al. ‘Panorama of the distal myopathies’. In: *Acta Myologica* 39.4 (Dec. 2020), p. 245. ISSN: 1128-2460.

[561] Bernhard Schölkopf et al. *Towards Causal Representation Learning*. Feb. 2021.

[562] Gesina Schwalbe and Bettina Finzel. ‘A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts’. In: *Data Mining and Knowledge Discovery* 38.5 (Jan. 2023), pp. 3043–3101. ISSN: 1573-756X.

[563] Fabrizio Sebastiani. ‘Machine Learning in Automated Text Categorization’. In: *ACM Comput. Surv.* 34.1 (Mar. 2002), pp. 1–47. ISSN: 0360-0300.

[564] Andrew D Selbst and Julia Powles. ‘Meaningful information and the right to explanation’. In: *International Data Privacy Law* 7.4 (Dec. 2017), pp. 233–242. ISSN: 2044-3994.

[565] Ramprasaath R Selvaraju et al. ‘Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization’. en. In: *Int. J. Comput. Vis.* 128.2 (Feb. 2020), pp. 336–359. ISSN: 0920-5691.

[566] Ramprasaath R Selvaraju et al. ‘Grad-cam: Visual explanations from deep networks via gradient-based localization’. In: *Proceedings of the IEEE international conference on computer vision*. IEEE, Oct. 2017, pp. 618–626.

[567] Rico Sennrich, Barry Haddow and Alexandra Birch. ‘Neural Machine Translation of Rare Words with Subword Units’. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. abs/1508.07909. Berlin, Germany: Association for Computational Linguistics, Aug. 2016, pp. 1715–1725.

[568] Septiarini, Anindita and Hamdani, Hamdani and Hatta, Heliza Rahmania and Anwar, Khoerul. ‘Automatic image segmentation of oil palm fruits by applying the contour-based approach’. In: *Scientia Horticulturae* 261 (Nov. 2019). Publisher: Elsevier BV, p. 108939. ISSN: 0304-4238.

[569] Sofia Serrano and Noah A. Smith. ‘Is Attention Interpretable?’ In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David R. Traum and Lluís Màrquez. event-place: Florence, Italy. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 2931–2951. (Visited on 31/05/2021).

[570] Lloyd S Shapley and Alvin E Roth. *The Shapley value: essays in honor of Lloyd S. Shapley*. en. Ed. by Lloyd S Shapley and Alvin E Roth. Cambridge [Cambridgeshire] ; New York: Cambridge University Press, 1988.

[571] Shikhar Sharma, Ryan Kiros and Ruslan Salakhutdinov. ‘Action Recognition using Visual Attention’. In: *arXiv:1511.04119 [cs]* (Feb. 2016). arXiv: 1511.04119. (Visited on 31/05/2021).

[572] Jiaming Shen et al. ‘TaxoClass: Hierarchical Multi-Label Text Classification Using Only Class Names’. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*. Ed. by Kristina Toutanova et al. Association for Computational Linguistics, June 2021, pp. 4239–4249.

[573] Shewfelt, R L and Thai, C N and Davis, J W. ‘Prediction of changes in color of tomatoes during ripening at different constant temperatures’. In: *Journal of Food Science* 53.5 (Sept. 1988). Publisher: Wiley, pp. 1433–1437. ISSN: 0022-1147.

[574] Lilong Shi and Brian Funt. *”Re-processed Version of the Gehler Color Constancy Dataset of 568 Images”*.

[575] Xingjian SHI et al. ‘Convolutional LSTM Network: A Machine Learning Approach for Precipitation Nowcasting’. In: *Advances in Neural Information Processing Systems*. Ed. by C. Cortes et al. Vol. 28. Nips’15. Montreal, Canada: Curran Associates, Inc., June 2015, pp. 802–810.

[576] Shiddiq, Minarni and Herman, Herman and Arief, Dodi Sofyan and Fitra, Edy and Husein, Ikhsan Rahman and Ningsih, Sinta Afria. ‘Wavelength selection of multispectral imaging for oil palm fresh fruit ripeness classification’. In: *Applied Optics* 61.17 (June 2022). Publisher: Optica Publishing Group, pp. 5289–5298. ISSN: 2155-3165. (Visited on 21/02/2023).

[577] Ben Shneiderman. *145th NOTE on Human-Centered AI*. `https://groups.google.com/g/human-centered-ai/c/NbPxZ3Hy6D8`. Accessed: 2025-11-05. Sept. 2025.

[578] Avanti Shrikumar, Peyton Greenside and Anshul Kundaje. ‘Learning Important Features Through Propagating Activation Differences’. In: *Proceedings of the 34th International Conference on Machine Learning*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. Proceedings of Machine Learning Research. Sydney, NSW, Australia: PMLR, May 2017, pp. 3145–3153.

[579] David I Shuman et al. ‘The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains’. In: *IEEE Signal Processing Magazine* 30.3 (May 2013), pp. 83–98. ISSN: 1053-5888.

[580] R. Shwartz-Ziv and A. Armon. ‘Tabular data: Deep learning is not all you need’. In: *arXiv preprint arXiv:2106.03253* 81 (June 2021), pp. 84–90. ISSN: 1566-2535.

[581] Laurent Sifre and Stéphane Mallat. ‘Rigid-Motion Scattering for Texture Classification’. In: *CoRR* abs/1403.1687 (Mar. 2014). ISSN: 2331-8422. arXiv: `1403.1687`.

[582] Silalahi, D D and Reano, C E and Lansigan, F P and Panopio, R G and Bantayan, N C. ‘Using genetic algorithm neural network on near infrared spectral data for ripeness grading of oil palm (elaeis guineensis jacq.) fresh fruit’. In: *Information Processing in Agriculture* 3.4 (Dec. 2016), pp. 252–261. ISSN: 2214-3173.

[583] Carlos N. Silla and Alex A. Freitas. ‘A survey of hierarchical classification across different application domains’. In: *Data Mining and Knowledge Discovery* 22.1 (Jan. 2011), pp. 31–72. ISSN: 1573-756X.

[584] Luan V. M. da Silva and Ricardo Cerri. ‘Feature Selection for Hierarchical Multi-label Classification’. In: *Advances in Intelligent Data Analysis XIX*. Ed. by Pedro Henriques Abreu et al. Vol. 12695. Cham: Springer International Publishing, 2021, pp. 196–208. ISBN: 978-3-030-74251-5.

[585] Jaspreet Singh and Avishek Anand. ‘EXS: Explainable Search Using Local Model Agnostic Interpretability’. en. In: *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. Ed. by J. Shane Culpepper et al. Melbourne VIC Australia: ACM, Jan. 2019, pp. 770–773. ISBN: 978-1-4503-5940-5. (Visited on 25/10/2021).

[586] V. Sitzmann et al. ‘Implicit neural representations with periodic activation functions’. In: *Advances in Neural Information Processing Systems*. Ed. by Hugo Larochelle et al. Vol. 33. Cornell University, June 2020, pp. 7462–7473.

[587] Junru Song, Feifei Wang and Yang Yang. ‘Peer-Label Assisted Hierarchical Text Classification’. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 3747–3758.

[588] Sijie Song et al. ‘An end-to-end spatio-temporal attention model for human action recognition from skeleton data’. In: *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*. Ed. by Satinder P. Singh and Shaul Markovitch. Vol. 31. Aaai’17 1. San Francisco, California, USA: AAAI Press, Feb. 2017, pp. 4263–4270. (Visited on 31/05/2021).

[589] Yong Song et al. ‘Hierarchical Multi-label Text Classification based on a Matrix Factorization and Recursive-Attention Approach’. In: *2022 7th International Conference on Big Data Analytics (ICBDA)*. Vol. 2007. IEEE, Mar. 2022, pp. 170–176.

[590] Song, Jun and Deng, Weimin and Beaudry, Randolph M and Armstrong, Paul R. ‘Changes in chlorophyll fluorescence of apple fruit during maturation, ripening, and senescence’. In: *HortScience* 32.5 (Aug. 1997). Publisher: American Society for Horticultural Science, pp. 891–896. ISSN: 0018-5345, 2327-9834.

[591] Thomas Soroski et al. ‘Differentiating memory clinic patients and healthy volunteers using machine-learning analysis of speech and eye movements during a reading task’. In: *Alzheimer’s & Dementia* 17.S6 (Dec. 2021), e055717. ISSN: 1552-5260. eprint: `https://alz-journals.onlinelibrary.wiley.com/doi/pdf/10.1002/alz.055717`.

[592] Majd Soud, Waltteri Nuutinen and Grischa Liebel. ‘Sóley: Automated detection of logic vulnerabilities in Ethereum smart contracts using large language models’. In: *Journal of Systems and Software* 226 (Mar. 2025), p. 112406. ISSN: 0164-1212.

[593] Robyn Speer, Joshua Chin and Catherine Havasi. ‘ConceptNet 5.5: An Open Multilingual Graph of General Knowledge’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 31.1 (Feb. 2017). Ed. by Satinder P. Singh and Shaul Markovitch. ISSN: 2159-5399.

[594] Speirs, J. and Lee, E. and Brady, C. J. and Robertson, J. and McGlasson, W. B. ‘Endopolygalacturonase: Messenger RNA, Enzyme and Softening in the Ripening Fruit of a Range of Tomato Genotypes’. In: *Journal of Plant Physiology* 135.5 (Jan. 1990), pp. 576–582. ISSN: 0176-1617. (Visited on 21/02/2023).

[595] Sidharth SS, Keerthana AR, Anas KP et al. ‘Chebyshev polynomial-based kolmogorov-arnold networks: An efficient architecture for nonlinear function approximation’. In: *arXiv preprint arXiv:2405.07200* (May 2024).

[596] Roger A. Stein, Patrícia Augustin Jaques and João Francisco Valiati. ‘An analysis of hierarchical text classification using word embeddings’. In: *Inf. Sci.* 471 (Jan. 2019), pp. 216–232. ISSN: 0020-0255.

[597] Tomaž Stepišnik and Dragi Kocev. ‘Hyperbolic Embeddings for Hierarchical Multi-label Classification’. In: *Foundations of Intelligent Systems*. Ed. by Denis Helic et al. Vol. 12117. Cham: Springer International Publishing, 2020, pp. 66–76. ISBN: 978-3-030-59491-6.

[598] Jianlin Su et al. ‘ZLPR: A Novel Loss for Multi-label Classification’. In: *ArXiv* abs/2208.02955 (Aug. 2022). ISSN: 2331-8422.

[599] Su, Zhenzhu and Zhang, Chu and Yan, Tianying and Zhu, Jianan and Zeng, Yulan and Lu, Xuanjun and Gao, Pan and Feng, Lei and He, Linhai and Fan, Lihui. ‘Application of Hyperspectral Imaging for Maturity and Soluble Solids Content Determination of Strawberry With Deep Learning Approaches’. In: *Frontiers in Plant Science* 12 (Sept. 2021), p. 736334. ISSN: 1664-462X. (Visited on 21/02/2023).

[600] Vivek Subbiah. ‘The next generation of evidence-based medicine’. en. In: *Nat. Med.* 29.1 (Jan. 2023), pp. 49–58. ISSN: 1078-8956.

[601] Suharjito and Elwirehardja, Gregorius Natanael and Prayoga, Jonathan Sebastian. ‘Oil palm fresh fruit bunch ripeness classification on mobile devices using deep learning approaches’. In: *Computers and Electronics in Agriculture* 188 (Sept. 2021). Publisher: Elsevier BV, p. 106359. ISSN: 0168-1699.

[602] A. Sun et al. ‘Blocking reduction strategies in hierarchical text classification’. In: *IEEE Transactions on Knowledge and Data Engineering* 16.10 (Oct. 2004), pp. 1305–1308. ISSN: 1041-4347.

[603] Aixin Sun and Ee-Peng Lim. ‘Hierarchical Text Classification and Evaluation’. In: *Proceedings of the 2001 IEEE International Conference on Data Mining*. Ed. by Nick Cercone, Tsau Young Lin and Xindong Wu. ICDM ’01. USA: IEEE Computer Society, Nov. 2001, pp. 521–528. ISBN: 0769511198.

[604] Aixin Sun, Ee-Peng Lim and Wee-Keong Ng. ‘Hierarchical Text Classification Methods and Their Specification’. In: *Cooperative Internet Computing*. Boston, MA: Springer US, 2003. Chap. 14, pp. 236–256. ISBN: 978-1-4615-0435-1.

[605] Xiaofei Sun et al. ‘Text Classification via Large Language Models’. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8990–9005.

[606] Sun, Chi and Qiu, Xipeng and Xu, Yige and Huang, Xuanjing. ‘How to Fine-Tune BERT for Text Classification?’ In: *Chinese Computational Linguistics*. Ed. by Maosong Sun et al. Vol. 11856. Lecture Notes in Computer Science. Cham: Springer International Publishing, May 2019, pp. 194–206. ISBN: 978-3-030-32381-3.

[607] Mukund Sundararajan, Ankur Taly and Qiqi Yan. 'Axiomatic attribution for deep networks'. In: *Proceedings of the 34th International Conference on Machine Learning - Volume 70*. Ed. by Doina Precup and Yee Whye Teh. Vol. 70. ICML'17. Sydney, NSW, Australia: JMLR.org, Mar. 2017, pp. 3319–3328.

[608] Sural, S and Qian, Gang and Pramanik, S. 'Segmentation and histogram generation using the HSV color space for image retrieval'. In: *Proceedings. International Conference on Image Processing*. Vol. 2. event-place: Rochester, NY, USA. IEEE, June 2003, pp. II–589.

[609] Ilya Sutskever, Oriol Vinyals and Quoc V. Le. 'Sequence to Sequence Learning with Neural Networks'. In: *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*. Ed. by Zoubin Ghahramani et al. Vol. abs/1409.3215. NIPS'14. Montreal, Canada: MIT Press, Sept. 2014, pp. 3104–3112.

[610] W C Swabey. 'The meaning of meaning: A study of the influence of language upon thought and of the science of symbolism'. In: *Philos. Rev.* 33.2 (Mar. 1924), pp. 222–223. ISSN: 0031-8108.

[611] Péter Szilágyi. *r/ethereum - How to PWN FoMo3D, a beginners guide*. `http://web.archive.org/web/20190105222409/https://www.reddit.com/r/ethereum/comments/916xni/how_to_pwn_fomo3d_a_beginners_guide/`. [Accessed 30-05-2025]. 2019.

[612] Kai Sheng Tai, Richard Socher and Christopher D. Manning. 'Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks'. In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. abs/1503.00075. Beijing, China: Association for Computational Linguistics, July 2015, pp. 1556–1566.

[613] Tamura, Hideyuki and Mori, Shunji and Yamawaki, Takashi. 'Textural features corresponding to visual perception'. In: *IEEE Transactions on Systems, Man, and Cybernetics* 8.6 (June 1978). Publisher: Institute of Electrical and Electronics Engineers (IEEE), pp. 460–473. ISSN: 0018-9472.

[614] Hirotaka Tanaka et al. 'Document Classification by Word Embeddings of BERT'. In: *16th International Conference of the Pacific Association for Computational Linguistics, PACLING 2019*. Ed. by Le-Minh Nguyen et al. Vol. 1215. Hanoi, Vietnam: Springer Singapore, Oct. 2019, pp. 145–154.

[615] Teng, L and Cheng, Z and Chen, X and Lai, L. 'Study on simulation models of tomato fruit quality related to cultivation environmental factors'. In: *Acta Ecologica Sinica* 32.2 (May 2012), pp. 111–116. ISSN: 1000-0933.

[616] Ian Tenney et al. 'The Language Interpretability Tool: Extensible, Interactive Visualizations and Analysis for NLP Models'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Online: Association for Computational Linguistics, Oct. 2020, pp. 107–118.

[617] Gabriel Terejanu et al. 'Explainable deep modeling of tabular data using tablegraphnet'. In: *arXiv preprint arXiv:2002.05205* abs/2002.05205 (Feb. 2020). ISSN: 2331-8422.

[618] R. Tibshirani. 'Regression shrinkage and selection via the lasso'. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (Jan. 1996), pp. 267–288. ISSN: 1369-7412.

[619] T. Tieleman and G. Hinton. *Lecture 6.5—RmsProp: Divide the gradient by a running average of its recent magnitude*. COURSERA: Neural Networks for Machine Learning. 2012.

[620] Tomana, T and Utsunomiya, N and Kataoka, I. 'The effect of environmental temperatures on fruit ripening on the tree'. In: *Journal of the Japanese Society for Hoticultural Science* 48.3 (1979), pp. 261–266. ISSN: 0013-7626.

[621] Sana Tonekaboni et al. 'What Clinicians Want: Contextualizing Explainable Machine Learning for Clinical End Use'. en. In: *Proceedings of the 4th Machine Learning for Healthcare Conference*. Ed. by Finale Doshi-Velez et al. Vol. abs/1905.05134. ISSN: 2640-3498. PMLR, Oct. 2019, pp. 359–380. (Visited on 29/05/2024).

[622] Eric Topol. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. en. Hachette UK, Mar. 2019.

[623] Christof Ferreira Torres, Julian Schütte and Radu State. 'Osiris: Hunting for integer bugs in ethereum smart contracts'. In: *Proceedings of the 34th Annual Computer Security Applications Conference*. ACM, Dec. 2018, pp. 664–676.

[624] Christof Ferreira Torres et al. 'Confuzzius: A data dependency-aware hybrid fuzzer for smart contracts'. In: *2021 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE, Sept. 2021, pp. 103–119.

[625] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. Feb. 2023. arXiv: `2302.13971 [cs.CL]`.

[626] L. N. Trefethen. *Approximation Theory and Approximation Practice*. SIAM, Jan. 2013.

[627] Shang-Chi Tsai, Chao-Wei Huang and Yun-Nung Chen. 'Modeling Diagnostic Label Correlation for Automatic ICD Coding'. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kristina Toutanova et al. Online: Association for Computational Linguistics, June 2021, pp. 4043–4052.

[628] Petar Tsankov et al. 'Securify: Practical security analysis of smart contracts'. In: *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*. Ed. by David Lie et al. ACM, June 2018, pp. 67–82.

[629] Ukirade, N S. 'Color grading system for evaluating tomato maturity'. In: *International Journal of Research in Management, Science & Technology* (2014). Pages: 41–45 Volume: 2.

[630] Ihsan Ullah et al. 'Explaining deep learning models for tabular data using layer-wise relevance propagation'. In: *Applied Sciences* 12.1 (Dec. 2021), p. 136. ISSN: 2076-3417.

[631] Uwadaira, Yasuhiro and Sekiyama, Yasuyo and Ikehata, Akifumi. 'An examination of the principle of non-destructive flesh firmness measurement of peach fruit by using VIS-NIR spectroscopy'. In: *Heliyon* 4.2 (Feb. 2018), e00531. ISSN: 2405-8440.

[632] V. Prasanna and R. N. Tharanathan. ‘Fruit ripening phenomena–an overview’. In: *Critical Reviews in Food Science and Nutrition* 47.1 (Jan. 2007), pp. 1–19. ISSN: 1040-8398.

[633] Bas HM Van der Velden et al. ‘Explainable artificial intelligence (XAI) in deep learning-based medical image analysis’. In: *Medical Image Analysis* 79 (July 2022), p. 102470. ISSN: 1361-8415.

[634] William van Melle. ‘MYCIN: a knowledge-based consultation program for infectious disease diagnosis’. In: *International Journal of Man-Machine Studies* 10.3 (May 1978), pp. 313–322. ISSN: 0020-7373.

[635] Ashish Vaswani et al. ‘Attention is All you Need’. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. NIPS’17. Long Beach, California, USA: Curran Associates, Inc., June 2017, pp. 6000–6010. ISBN: 9781510860964.

[636] Ashwin Vaswani et al. ‘All Mistakes Are Not Equal: Comprehensive Hierarchy Aware Multi-label Predictions (CHAMP)’. In: *arXiv (preprint)* abs/2206.08653 (June 2022). ISSN: 2331-8422.

[637] Henry Veatch. ‘Carl G. Hempel Aspects of scientific explanation and other essays in the philosophy of science. New York: The Free Press, 1965. 505 pp.’ In: *Philosophy of Science* 37.2 (June 1970), pp. 312–314. ISSN: 1539-767X.

[638] Petar Veličković et al. ‘Graph Attention Networks’. In: *International Conference on Learning Representations*. Cornell University, Feb. 2018.

[639] Celine Vens et al. ‘Decision trees for hierarchical multi-label classification’. In: *Machine Learning* 73.2 (Aug. 2008), p. 185. ISSN: 1573-0565.

[640] José Verdú-Díaz et al. ‘Accuracy of a machine learning muscle MRI-based tool for the diagnosis of muscular dystrophies’. In: *Neurology* 94.10 (Feb. 2020), e1094–e1102. ISSN: 0028-3878.

[641] Manisha Verma and Debasis Ganguly. ‘LIRME: Locally Interpretable Ranking Model Explanation’. In: *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ed. by Benjamin Piwowarski et al. SIGIR’19. New York, NY, USA: Association for Computing Machinery, July 2019, pp. 1281–1284. ISBN: 978-1-4503-6172-9. (Visited on 04/11/2021).

[642] Giulia Vilone and Luca Longo. *Explainable Artificial Intelligence: a Systematic Review*. May 2020. arXiv: 2006.00093 [cs.AI].

[643] Lalitkumar K Vora et al. ‘Artificial intelligence in pharmaceutical technology and drug delivery design’. In: *Pharmaceutics* 15.7 (July 2023), p. 1916. ISSN: 1999-4923.

[644] Wietse de Vries, Andreas van Cranenburgh and Malvina Nissim. ‘What’s so special about BERT’s layers? A closer look at the NLP pipeline in monolingual and multilingual models’. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Ed. by Trevor Cohn, Yulan He and Yang Liu. Vol. abs/2004.06499. Online: Association for Computational Linguistics, Nov. 2020, pp. 4339–4350.

[645] Shiyun Wa, Xinai Lu and Minjuan Wang. ‘Stable and Interpretable Deep Learning for Tabular Data: Introducing InterpreTabNet with the Novel InterpreStability Metric’. In: *arXiv preprint arXiv:2310.02870* abs/2310.02870 (Oct. 2023). ISSN: 2331-8422.

[646] Sandra Wachter, Brent Mittelstadt and Chris Russell. ‘Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR’. In: *SSRN Electronic Journal* (Nov. 2017). ISSN: 1556-5068.

[647] Yasmen Wahba, Nazim H. Madhavji and John Steinbacher. ‘A Hybrid Continual Learning Approach for Efficient Hierarchical Classification of IT Support Tickets in the Presence of Class Overlap’. In: *2023 IEEE International Conference on Industrial Technology (ICIT)*. IEEE, May 2023, pp. 1–6.

[648] Walsh, K B and Blasco, J and Zude-Sasse, M and Sun, X. ‘Visible-NIR ‘point’ spectroscopy in postharvest fruit and vegetable assessment: The science behind three decades of commercial use’. In: *Postharvest Biology and Technology* 168 (Oct. 2020), p. 111246. ISSN: 0925-5214.

[649] Alex Wang et al. ‘GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding’. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Ed. by Tal Linzen, Grzegorz Chrupala and Afra Alishahi. Brussels, Belgium: Association for Computational Linguistics, May 2018, pp. 353–355.

[650] Alex Wang et al. ‘SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems’. In: *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Ed. by Hanna M. Wallach et al. Vol. abs/1905.00537. Red Hook, NY, USA: Curran Associates Inc., May 2019, pp. 3266–3280.

[651] Alex X Wang et al. ‘Challenges and opportunities of generative models on tabular data’. In: *Applied Soft Computing* 166 (Nov. 2024), p. 112223. ISSN: 1568-4946.

[652] Boyan Wang et al. ‘Cognitive structure learning model for hierarchical multi-label text classification’. In: *Knowl. Based Syst.* 218 (May 2021), p. 106876. ISSN: 0950-7051.

[653] Daixin Wang, Peng Cui and Wenwu Zhu. ‘Structural Deep Network Embedding’. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Ed. by Balaji Krishnapuram et al. KDD ’16. San Francisco, California, USA: Association for Computing Machinery, Aug. 2016, pp. 1225–1234. ISBN: 9781450342322.

[654] Kuansan Wang et al. ‘Microsoft Academic Graph: When experts are not enough’. In: *Quantitative Science Studies* 1.1 (Feb. 2020), pp. 396–413. ISSN: 2641-3337. eprint: https://direct.mit.edu/qss/article-pdf/1/1/396/1760880/qss_a_00021.pdf.

[655] Ning Wang et al. ‘Video-Based Illumination Estimation’. In: *Computational Color Imaging*. Ed. by Raimondo Schettini, Shoji Tominaga and Alain Trémeau. Vol. 6626. Berlin, Heidelberg: Springer Berlin Heidelberg, May 2011, pp. 188–198. ISBN: 978-3-642-20404-3.

[656] Qingyuan Wang et al. ‘Justifying the Importance of Color Cues in Object Detection: A Case Study on Pedestrian’. In: *The Era of Interactive Media* 1 (Oct. 2013), pp. 387–397.

[657] Xin Wang and Leifeng Guo. ‘Multi-Label Classification of Chinese Rural Poverty Governance Texts Based on XLNet and Bi-LSTM Fused Hierarchical Attention Mechanism’. In: *Applied Sciences* 13.13 (June 2023), p. 7377. ISSN: 2076-3417.

[658] Xuepeng Wang et al. ‘Concept-Based Label Embedding via Dynamic Routing for Hierarchical Text Classification’. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. Ed. by Chengqing Zong et al. Association for Computational Linguistics, 2021, pp. 5010–5019.

[659] Yuan Wang et al. ‘Exploiting Dynamic and Fine-grained Semantic Scope for Extreme Multi-label Text Classification’. In: *Natural Language Processing and Chinese Computing*. Ed. by Wei Lu et al. Vol. 13552. Cham: Springer Nature Switzerland, May 2022, pp. 85–97. ISBN: 978-3-031-17189-5.

[660] Yue Wang et al. ‘Towards Better Hierarchical Text Classification with Data Generation’. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 7722–7739.

[661] Yujing Wang et al. ‘TextNAS: A Neural Architecture Search Space Tailored for Text Representation’. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.05 (Apr. 2020), pp. 9242–9249. ISSN: 2159-5399.

[662] Zeyu Wang. ‘CausalBench: A Comprehensive Benchmark for Evaluating Causal Reasoning Capabilities of Large Language Models’. In: *Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10)*. Ed. by Kam-Fai Wong et al. Bangkok, Thailand: Association for Computational Linguistics, Aug. 2024, pp. 143–151.

[663] Zhen Wang et al. ‘Knowledge graph embedding by translating on hyperplanes’. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. Ed. by Carla E. Brodley and Peter Stone. Vol. 28. AAAI’14 1. Québec City, Québec, Canada: AAAI Press, June 2014, pp. 1112–1119.

[664] Zhou Wang et al. ‘Image quality assessment: from error visibility to structural similarity’. In: *IEEE Transactions on Image Processing* 13.4 (May 2004), pp. 600–612. ISSN: 1057-7149.

[665] Zihan Wang et al. ‘HPT: Hierarchy-aware Prompt Tuning for Hierarchical Text Classification’. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Ed. by Yoav Goldberg, Zornitsa Kozareva and Yue Zhang. Vol. abs/2204.13413. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022, pp. 3740–3751.

[666] Zihan Wang et al. ‘Incorporating Hierarchy into Text Encoder: a Contrastive Learning Approach for Hierarchical Text Classification’. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Smaranda Muresan, Preslav Nakov and Aline Villavicencio. Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 7109–7119.

[667] Wang, Hailong and Peng, Jiyu and Xie, Chuanqi and Bao, Yidan and He, Yong. ‘Fruit quality evaluation using spectroscopy technology: a review’. In: *Sensors* 15.5 (May 2015), pp. 11889–11927. ISSN: 1424-8220.

[668] Wang, Shuang and Huang, Min and Zhu, Qibing. ‘Model fusion for prediction of apple firmness using hyperspectral scattering image’. In: *Computers and Electronics in Agriculture* 80 (Jan. 2012). Publisher: Elsevier BV, pp. 1–7. ISSN: 0168-1699.

[669] Niyaz Ahmad Wani et al. ‘Explainable AI-driven IoMT fusion: Unravelling techniques, opportunities, and challenges with Explainable AI in healthcare’. In: *Information Fusion* 110 (Oct. 2024), p. 102472. ISSN: 1566-2535.

[670] Wankhade, M and Hore, U W. ‘A Survey on Fruit Ripeness Classification Based On Image Processing with Machine Learning’. In: *International Journal of Advanced Research in Science, Communication and Technology* (May 2021), pp. 73–78. ISSN: 2581-9429.

[671] Patrick Weber, K Valerie Carl and Oliver Hinz. ‘Applications of explainable artificial intelligence in finance–a systematic review of finance, information systems, and computer science literature’. In: *Management Review Quarterly* 74.2 (June 2024), pp. 867–907. ISSN: 2198-1620.

[672] Jonatas Wehrmann, Ricardo Cerri and Rodrigo Barros. ‘Hierarchical Multi-Label Classification Networks’. In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, July 2018, pp. 5075–5084.

[673] Jason Wei et al. ‘Chain-of-thought prompting elicits reasoning in large language models’. In: *Proceedings of the 36th International Conference on Neural Information Processing Systems*. Ed. by Sanmi Koyejo et al. Vol. abs/2201.11903. NIPS ’22. New Orleans, LA, USA: Curran Associates Inc., Jan. 2022. ISBN: 9781713871088.

[674] Wei, Xuan and Liu, Fei and Qiu, Zhengjun and Shao, Yongni and He, Yong. ‘Ripeness classification of astringent persimmon using hyperspectral imaging technique’. In: *Food and Bioprocess Technology* 7.5 (May 2014). Publisher: Springer Science and Business Media LLC, pp. 1371–1380. ISSN: 1935-5130.

[675] Sarah Wiegreffe and Yuval Pinter. ‘Attention is not not Explanation’. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Ed. by Kentaro Inui et al. event-place: Hong Kong, China. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 11–20. (Visited on 31/05/2021).

[676] Jenna Wiens et al. ‘A Study on Bias and Fairness in Machine Learning’. In: *Nat. Med.* (2019).

[677] Wismadi, I M and Khrisne, D C and Others. ‘Detecting the ripeness of harvest-ready dragon fruit using smaller VGGNet-like network’. In: *Journal of Electrical and Electronics Engineering Australia* 3.2 (Jan. 2020). Number: 2, p. 35. ISSN: 2549-8304. (Visited on 21/02/2023).

[678] Thomas Wolf et al. 'Transformers: State-of-the-Art Natural Language Processing'. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Ed. by Qun Liu and David Schlangen. Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[679] Bichen Wu et al. *Visual Transformers: Token-based Image Representation and Processing for Computer Vision*. June 2020. arXiv: 2006.03677 [cs.CV].

[680] Chunyang Wu et al. 'Improving Interpretability and Regularization in Deep Learning'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26.2 (Feb. 2018), pp. 256–265. ISSN: 2329-9290.

[681] Hongjun Wu et al. 'Peculiar: Smart contract vulnerability detection based on crucial data flow graph and pre-training techniques'. In: *2021 IEEE 32nd International Symposium on Software Reliability Engineering (ISSRE)*. Ed. by Zhi Jin et al. IEEE, Oct. 2021, pp. 378–389.

[682] Tongshuang Wu et al. 'Errudite: Scalable, Reproducible, and Testable Error Analysis'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Ed. by Anna Korhonen, David R. Traum and Lluís Màrquez. Florence, Italy: Association for Computational Linguistics, July 2019, pp. 747–763.

[683] Yuqian Wu, Hengyi Luo and Raymond ST Lee. 'Deep feature embedding for tabular data'. In: *arXiv preprint arXiv:2408.17162* abs/2408.17162 (Aug. 2024). ISSN: 2331-8422.

[684] Zichao Wu et al. 'Syntax-Aware Retrieval for Generation-based Code Synthesis'. In: *Proceedings of the 38th IEEE/ACM International Conference on Automated Software Engineering*. 2023, pp. 121–132.

[685] Zonghan Wu et al. 'A Comprehensive Survey on Graph Neural Networks'. In: *IEEE Transactions on Neural Networks and Learning Systems* 32.1 (Jan. 2021), pp. 4–24. ISSN: 2162-237X.

[686] Dominik Wunderlich et al. 'On the Privacy & Utility Trade-Off in Differentially Private Hierarchical Text Classification'. In: *Applied Sciences* 12.21 (Nov. 2022), p. 11177. ISSN: 2076-3417.

[687] Magdalena Wysocka et al. 'A systematic review of biologically-informed deep learning models for cancer: fundamental trends for encoding and interpreting oncology data'. en. In: *BMC Bioinformatics* 24.1 (May 2023), p. 198. ISSN: 1471-2105.

[688] Edgar Xi, Selina Bing and Yang Jin. 'Capsule Network Performance on Complex Data'. In: *arXiv (preprint)* abs/1712.03480 (Dec. 2017). ISSN: 2331-8422.

[689] Huiru Xiao, Xin Liu and Yangqiu Song. 'Efficient Path Prediction for Semi-Supervised and Weakly Supervised Hierarchical Text Classification'. In: *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*. Ed. by Ling Liu et al. ACM, May 2019, pp. 3370–3376.

[690] Jincheng Xu and Qingfeng Du. 'Learning neural networks for text classification by exploiting label relations'. In: *Multimedia Tools and Applications* 79.31 (Aug. 2020), pp. 22551–22567. ISSN: 1573-7721.

[691] Jun Xu et al. 'Learning Multimodal Attention LSTM Networks for Video Captioning'. In: *Proceedings of the 25th ACM international conference on Multimedia*. Ed. by Qiong Liu et al. Mm '17. New York, NY, USA: Association for Computing Machinery, Oct. 2017, pp. 537–545. ISBN: 978-1-4503-4906-2. (Visited on 31/05/2021).

[692] Kelvin Xu et al. 'Show, attend and tell: neural image caption generation with visual attention'. In: *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*. Ed. by Francis R. Bach and David M. Blei. Vol. 37. Icml'15. Lille, France: JMLR.org, July 2015, pp. 2048–2057. (Visited on 31/05/2021).

[693] Linli Xu et al. 'Hierarchical Multi-label Text Classification with Horizontal and Vertical Category Correlations'. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. Ed. by Marie-Francine Moens et al. Association for Computational Linguistics, 2021, pp. 2459–2468.

[694] Shuo Xu, Yan Li and Zheng Wang. 'Bayesian Multinomial Naïve Bayes Classifier to Text Classification'. In: *Advanced Multimedia and Ubiquitous Engineering*. Ed. by James Jong Hyuk Park, Shu-Ching Chen and Kim-Kwang Raymond Choo. Vol. 448. Springer Singapore, May 2017, pp. 347–352. ISBN: 978-981-10-5041-1.

[695] Yang Xu et al. *Tracking the Feature Dynamics in LLM Training: A Mechanistic Study*. Dec. 2024.

[696] Zhikang Xu et al. 'Hierarchical multilabel classification by exploiting label correlations'. In: *International Journal of Machine Learning and Cybernetics* 13.1 (Jan. 2022), pp. 115–131. ISSN: 1868-808X.

[697] Guangxu Xun et al. 'Correlation Networks for Extreme Multi-Label Text Classification'. In: *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. Ed. by Rajesh Gupta et al. KDD '20. Virtual Event, CA, USA: Association for Computing Machinery, Aug. 2020, pp. 1074–1082. ISBN: 9781450379984.

[698] Yam, Kit L. and Papadakis, Spyridon E. 'A simple digital imaging method for measuring and analyzing color of food surfaces'. In: *Journal of Food Engineering*. Applications of computer vision in the food industry 61.1 (Jan. 2004), pp. 137–142. ISSN: 0260-8774. (Visited on 21/02/2023).

[699] Yamada, Hisashi and Ohmura, Hirokazu and Arai, Chizuru and Terui, Makoto. 'Effect of Preharvest Fruit Temperature on Ripening, Sugars, and Watercore Occurrence in Apples'. In: *Journal of the American Society for Horticultural Science* 119.6 (Nov. 1994). Publisher: American Society for Horticultural Science Section: Journal of the American Society for Horticultural Science, pp. 1208–1214. ISSN: 2327-9788, 0003-1062. (Visited on 21/02/2023).

[700] Chenggang Yan et al. 'STAT: Spatial-Temporal Attention Mechanism for Video Captioning'. In: *IEEE Transactions on Multimedia* 22.1 (Jan. 2020). Conference Name: IEEE Transactions on Multimedia, pp. 229–241. ISSN: 1941-0077.

[701] Jingsong Yan et al. 'Does the Order Matter? A Random Generative Way to Learn Label Hierarchy for Hierarchical Text Classification'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2024), pp. 276–285. ISSN: 2329-9290.

[702] Jining Yan et al. ‘Temporal Convolutional Networks for the Advance Prediction of ENSO’. In: *Scientific Reports* 10.1 (May 2020), p. 8055. ISSN: 2045-2322.

[703] Boyang Yang et al. ‘A Hybrid RAG System with Comprehensive Enhancement on Complex Reasoning’. In: *arXiv preprint arXiv:2408.05141* (Aug. 2024). arXiv: 2408.05141 [cs.CL].

[704] Hao Yang et al. ‘STA-CNN: Convolutional Spatial-Temporal Attention Learning for Action Recognition’. In: *IEEE Transactions on Image Processing* 29 (May 2020). Conference Name: IEEE Transactions on Image Processing, pp. 5783–5793. ISSN: 1941-0042.

[705] Mei Yang et al. ‘A deep learning model for diagnosing dystrophinopathies on thigh muscle MRI images’. In: *BMC neurology* 21.1 (Jan. 2021), pp. 1–9. ISSN: 1471-2377.

[706] Pengcheng Yang et al. ‘SGM: Sequence Generation Model for Multi-label Classification’. In: *Proceedings of the 27th International Conference on Computational Linguistics*. Ed. by Emily M. Bender, Leon Derczynski and Pierre Isabelle. Santa Fe, New Mexico, USA: Association for Computational Linguistics, Aug. 2018, pp. 3915–3926.

[707] Xu Yang et al. ‘Causal Attention for Vision-Language Tasks’. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2021, pp. 9847–9857.

[708] Yi Yang et al. ‘Effective Seed-Guided Topic Labeling for Dataless Hierarchical Short Text Classification’. In: *Web Engineering - 21st International Conference, ICWE 2021, Biarritz, France, May 18-21, 2021, Proceedings*. Ed. by Marco Brambilla et al. Vol. 12706. Lecture Notes in Computer Science. Springer, 2021, pp. 271–285. ISBN: 978-3030742959.

[709] Yingxiang Yang et al. ‘Fourier Learning with Cyclical Data’. In: *International Conference on Machine Learning*. Ed. by Kamalika Chaudhuri et al. Vol. 162. PMLR. PMLR, 2022, pp. 25280–25301.

[710] Zhenyu Yang and Guojing Liu. ‘Hierarchical Sequence-to-Sequence Model for Multi-Label Text Classification’. In: *IEEE Access* 7 (Oct. 2019), pp. 153012–153020. ISSN: 2169-3536.

[711] Zichao Yang et al. ‘Hierarchical Attention Networks for Document Classification’. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Kevin Knight, Ani Nenkova and Owen Rambow. San Diego, California: Association for Computational Linguistics, June 2016, pp. 1480–1489. (Visited on 31/05/2021).

[712] Yang, Haiqing. ‘Remote sensing technique for predicting harvest time of tomatoes’. In: *Procedia Environmental Sciences* 10 (2011). Publisher: Elsevier BV, pp. 666–671. ISSN: 1878-0296.

[713] Yang, S F and Hoffman, N E. ‘Ethylene biosynthesis and its regulation in higher plants’. In: *Annual Review of Plant Physiology* 35.1 (June 1984), pp. 155–189. ISSN: 0066-4294.

[714] Liang Yao, Chengsheng Mao and Yuan Luo. ‘Graph Convolutional Networks for Text Classification’. In: *Proc. AAAI Conf. Artif. Intell.* 33.01 (July 2019), pp. 7370–7377. ISSN: 2159-5399.

[715] Ziyi Yao et al. ‘HITSZQ at SemEval-2023 Task 10: Category-aware Sexism Detection Model with Self-training Strategy’. In: *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Ed. by Atul Kr. Ojha et al. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 934–940.

[716] Chenchen Ye et al. ‘Beyond Text: Incorporating Metadata and Label Structure for Multi-Label Document Classification using Heterogeneous Graphs’. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Ed. by Marie-Francine Moens et al. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 3162–3171.

[717] Chengxuan Ying et al. ‘Do Transformers Really Perform Badly for Graph Representation?’ In: *Advances in Neural Information Processing Systems*. Ed. by M. Ranzato et al. Vol. 34. Curran Associates, Inc., Dec. 2021, pp. 28877–28888.

[718] J. Yoo and J. Kim. ‘Dichromatic Model Based Temporal Color Constancy for AC Light Sources’. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2019, pp. 12321–12330.

[719] Junyong You and Jari Korhonen. ‘Attention Boosted Deep Networks For Video Classification’. In: *2020 IEEE International Conference on Image Processing (ICIP)*. Issn: 2381-8549. IEEE, Oct. 2020, pp. 1761–1765.

[720] Chao Yu, Yi Shen and Yue Mao. ‘Constrained Sequence-to-Tree Generation for Hierarchical Text Classification’. In: *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Ed. by Enrique Amigó et al. SIGIR ’22. Madrid, Spain: Association for Computing Machinery, July 2022, pp. 1865–1869. ISBN: 9781450387323.

[721] Huanglin Yu et al. ‘Cascading Convolutional Color Constancy’. In: *Aaai*. Vol. 34. 07. Association for the Advancement of Artificial Intelligence (AAAI), May 2020, pp. 12725–12732.

[722] Jeffy Yu. *Retrieval Augmented Generation Integrated Large Language Models in Smart Contract Vulnerability Detection*. arXiv preprint. July 2024. arXiv: 2407.14838 [cs.SE].

[723] Simon Chi Lok Yu et al. ‘Instances and Labels: Hierarchy-aware Joint Supervised Contrastive Learning for Hierarchical Multi-Label Text Classification’. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino and Kalika Bali. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 8858–8875.

[724] Yipeng Yu et al. ‘Hierarchical Multilabel Text Classification via Multitask Learning’. In: *33rd IEEE International Conference on Tools with Artificial Intelligence, ICTAI 2021, Washington, DC, USA, November 1-3, 2021*. Vol. 3. IEEE, Nov. 2021, pp. 1138–1143.

[725] Jun Yuan, Jesse Vig and Nazneen Rajani. ‘ISEA: An Interactive Pipeline for Semantic Error Analysis of NLP Models’. In: *27th International Conference on Intelligent User Interfaces*. Ed. by Giulio Jacucci et al. IUI ’22. Helsinki, Finland: Association for Computing Machinery, Mar. 2022, pp. 878–888. ISBN: 9781450391443.

[726] Tian Yuan and Xueming Li. 'Full Convolutional Color Constancy with Attention'. en. In: *Image and Graphics Technologies and Applications*. Springer, Singapore, Sept. 2020, pp. 114–126. ISBN: 978-9813360327. (Visited on 31/05/2021).

[727] A. Zangari et al. 'Crossing the divide: Designing layers of explainability'. In: *Artificial Intelligence and Soft Computing*. Ed. by L. Rutkowski et al. Cham: Springer, 2025, pp. 253–265. ISBN: 978-3031843525. DOI: 10.1007/978-3-031-84353-2_22. URL: https://doi.org/10.1007/978-3-031-84353-2_22.

[728] A. Zangari et al. 'Hierarchical text classification and its foundations: A review'. In: *Electronics* 13.7 (Mar. 2024), p. 1199. ISSN: 2079-9292. DOI: 10.3390/electronics13071199. URL: https://www.mdpi.com/2079-9292/13/7/1199/pdf?version=1711370387.

[729] Alessandro Zangari et al. *[dataset] Hierarchical Text Classification corpora (v.1), Zenodo.org*. Version 1.0. Nov. 2022.

[730] Matthew D Zeiler and Rob Fergus. 'Visualizing and understanding convolutional networks'. In: *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*. Ed. by David J. Fleet et al. Vol. 8689. Springer. Springer Science+Business Media, 2014, pp. 818–833. ISBN: 978-3319105895.

[731] Min-Ling Zhang and Zhi-Hua Zhou. 'A Review on Multi-Label Learning Algorithms'. In: *IEEE Transactions on Knowledge and Data Engineering* 26.8 (Aug. 2014), pp. 1819–1837. ISSN: 1041-4347.

[732] Quanshi Zhang, Ying Nian Wu and Song-Chun Zhu. 'Interpretable Convolutional Neural Networks'. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, June 2018, pp. 8827–8836.

[733] Xinyi Zhang et al. 'LA-HCN: Label-based Attention for Hierarchical Multi-label Text Classification Neural Network'. In: *Expert Syst. Appl.* 187 (Jan. 2022), p. 115922. ISSN: 0957-4174.

[734] Ye Zhang and Byron C. Wallace. 'A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification'. In: *Proceedings of the Eighth International Joint Conference on Natural Language Processing (IJCNLP)*. Ed. by Greg Kondrak and Taro Watanabe. Vol. 1. Taipei, Taiwan: Asian Federation of Natural Language Processing, Nov. 2017, pp. 253–263.

[735] Yilang Zhang et al. 'ADCC: An Effective and Intelligent Attention Dense Color Constancy System for Studying Images in Smart Cities'. In: *arXiv:1911.07163 [cs]* (Oct. 2020). arXiv: 1911.07163. (Visited on 31/05/2021).

[736] Yu Zhang et al. 'MATCH: Metadata-Aware Text Classification in A Large Hierarchy'. In: *Proceedings of the Web Conference 2021*. Ed. by Jure Leskovec et al. WWW '21. Ljubljana, Slovenia: Association for Computing Machinery, May 2021, pp. 3246–3257. ISBN: 9781450383127.

[737] Fei Zhao et al. 'Label-Correction Capsule Network for Hierarchical Text Classification'. In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 31 (2023), pp. 2158–2168. ISSN: 2329-9290.

[738] Haiyan Zhao et al. *Explainability for Large Language Models: A Survey*. Sept. 2023.

[739] Haiyan Zhao et al. 'Explainability for Large Language Models: A Survey'. In: *ACM Trans. Intell. Syst. Technol.* 15.2 (Feb. 2024), pp. 1–38. ISSN: 2157-6904.

[740] Rui Zhao et al. 'Hierarchical Multi-label Text Classification: Self-adaption Semantic Awareness Network Integrating Text Topic and Label Level Information'. In: *Knowledge Science, Engineering and Management*. Ed. by Han Qiu et al. Vol. 12816. Cham: Springer International Publishing, 2021, pp. 406–418. ISBN: 978-3-030-82147-0.

[741] Wei Zhao et al. 'Generative Multi-Task Learning for Text Classification'. In: *IEEE Access* 8 (2020), pp. 86380–86387. ISSN: 2169-3536.

[742] Zhou Zhao et al. 'Video Question Answering via Hierarchical Spatio-Temporal Attention Networks'. In: *Proceedings of the 26th International Joint Conference on Artificial Intelligence*. Ed. by Carles Sierra. IJCAI'17. Melbourne, Australia: AAAI Press, Aug. 2017, pp. 3518–3524. ISBN: 9780999241103.

[743] Zhao, Yuanshen and Gong, Liang and Huang, Yixiang and Liu, Chengliang. 'Robust Tomato Recognition for Robotic Harvesting Using Feature Images Fusion'. In: *Sensors* 16.2 (Jan. 2016), p. 173. ISSN: 1424-8220.

[744] Siqi Zheng et al. 'Label-Dividing Gated Graph Neural Network for Hierarchical Text Classification'. In: *2022 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2022, pp. 01–08.

[745] Zibin Zheng et al. 'Turn the Rudder: A Beacon of Reentrancy Detection for Smart Contracts on Ethereum'. In: *Proceedings of the 45th International Conference on Software Engineering*. ICSE '23. Melbourne, Victoria, Australia: IEEE Press, May 2023, pp. 295–306. ISBN: 9781665457019.

[746] Xiaoting Zhong et al. 'Explainable machine learning in materials science'. en. In: *npj Computational Materials* 8.1 (Sept. 2022), pp. 1–19. ISSN: 2057-3960.

[747] Bolei Zhou et al. 'Learning deep features for discriminative localization'. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. IEEE Computer Society, June 2016, pp. 2921–2929.

[748] Jie Zhou et al. 'Graph neural networks: A review of methods and applications'. In: *AI Open* 1 (2020), pp. 57–81. ISSN: 2666-6510.

[749] Jie Zhou et al. 'Hierarchy-Aware Global Model for Hierarchical Text Classification'. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*. Ed. by Dan Jurafsky et al. Association for Computational Linguistics, July 2020, pp. 1106–1117.

[750] He Zhu et al. 'HiTIN: Hierarchy-aware Tree Isomorphism Network for Hierarchical Text Classification'. In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Anna Rogers, Jordan Boyd-Graber and Naoaki Okazaki. Vol. abs/2305.15182. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 7809–7821.

[751] Huiming Zhu et al. ‘Patent Automatic Classification Based on Symmetric Hierarchical Convolution Neural Network’. In: *Symmetry* 12.2 (Jan. 2020), p. 186. ISSN: 2073-8994.

[752] Yitan Zhu et al. ‘Converting tabular data into images for deep learning with convolutional neural networks’. In: *Scientific reports* 11.1 (May 2021), p. 11325. ISSN: 2045-2322.

[753] Yongqing Zhu and Shuqiang Jiang. ‘Attention-based Densely Connected LSTM for Video Captioning’. In: *Proceedings of the 27th ACM International Conference on Multimedia*. Ed. by Laurent Amsaleg et al. Mm ’19. New York, NY, USA: Association for Computing Machinery, Oct. 2019, pp. 802–810. ISBN: 978-1-4503-6889-6. (Visited on 31/05/2021).

[754] Zhu, Qibing and Huang, Min and Zhao, Xin and Wang, Shuang. ‘Wavelength Selection of Hyperspectral Scattering Image Using New Semi-supervised Affinity Propagation for Prediction of Firmness and Soluble Solid Content in Apples’. In: *Food Analytical Methods* 6.1 (Feb. 2013), pp. 334–342. ISSN: 1936-976X. (Visited on 21/02/2023).

[755] Ziosi, V and Noferini, M and Fiori, G and Tadiello, A and Trainotti, L and Casadoro, G and Costa, G. ‘A new index based on vis spectroscopy to characterize the progression of ripening in peach fruit’. In: *Postharvest Biology and Technology* 49.3 (Sept. 2008). Publisher: Elsevier BV, pp. 319–329. ISSN: 0925-5214.

[756] Weilin Zou et al. ‘SmartDefi: A Multi-Aspect and Multi-Modal Neural Network Based Framework for Detecting Vulnerabilities in Smart Contracts’. In: *32nd USENIX Security Symposium (USENIX Security 23)*. 2023, pp. 555–572. ISBN: 978-1-939133-37-3.

Università
Ca'Foscari
Venezia

Borsa di dottorato cofinanziata con risorse dell'Unione europea-*NextGeneration EU*
Piano Nazionale di Ripresa e Resilienza Missione 4 – Componente 1 – Riforma 4.1 Riforma dei Dottorati

M4C2 – Investimento 3.3 - "Introduzione di dottorati innovativi che rispondono ai fabbisogni di innovazione delle imprese e promuovono l'assunzione dei ricercatori dalle imprese"

*CUP H73C22000340004*