Uncover to Discover: A Review of Explainable Artificial Intelligence for Knowledge Discovery

Bachelor's Project Thesis

Daniella Alves, s5492890, d.alves@student.rug.nl,
Supervisor: Dr. Jakob Schoeffer

Abstract: Explainable Artificial Intelligence (XAI) advocates for insight into models' reasoning through interpretable models and post-hoc explanation methods. With existing understanding of XAI used to justify, control, and improve, there is room to elaborate on how XAI has been utilised for the purpose of discovery. This study synthesises 42 research papers focusing on XAI for knowledge discovery, exploring the range of methods used. Furthermore, a taxonomy of discovery level is developed to analyse the findings and novelty potential of the examined studies. In doing so, this review assesses the landscape of XAI development, noting the ways in which XAI facilitates discovery, the potential risks of unguarded trust in XAI methods, and proposing to view discovery as a spectrum rather than a restricted definition.

1 Introduction

Machine Learning (ML) systems have become heavily integrated into diverse fields of research, leading towards greater dependence during high stakes decision-making processes. However, a system's lack of interpretability restricts the extent of its application, as it becomes increasingly challenging to comprehend the reasoning behind the conclusions of complex Artificial Intelligence (AI). Explainable Artificial Intelligence (XAI) advocates for the transparency that 'black box' methods lack, necessitating that insights gained or decisions-made based on AI findings are not limited by lack of understanding (Barredo et al., 2020).

Prior reviews conducted about XAI survey the overall landscape of XAI usage, often centred around organising current technology into taxonomies and developing definitions for keyword consensus. One such article by Barredo et al. (2020) examined XAI through the lens of two taxonomies: transparent and post-hoc methods. In this, they provided crucial clarification on XAI-related terms, such as interpretability and explainability, which tend to be conflated in existing literature. Their stance was that interpretability is a passive characteristic of a model regarding how easily it can be understood, whereas explainability is actively applied to intentionally provide insight (Barredo et al., 2020). Furthermore, they distinguished between models that are built to be interpretable in themselves and explainability methods that elaborate on the predictions of already developed models. These distinctions are critical in developing an understanding of what types of explainable technology are being utilised and the different forms in which they may be constructed.

Additionally, examining these methods can lead the conversation towards why XAI is becoming a focus, generating a dialogue surrounding the diverse purposes of XAI. Notably, Adadi & Berrada (2018) categorised the purpose of XAI adoption into four sectors, namely 'Explain to Justify', 'Explain to Control', 'Explain to Improve', and 'Explain to Discover'. Where elements of control and improvement focus on individual factors to streamline system performance; justification and discovery explore the reasoning behind decisions made by the system, and ultimately what new knowledge those findings and underlying methods can reveal.

Particularly, the purpose of 'Explain to Discover' is under-defined, as there is a lack of expansion regarding what can be considered new knowledge and how specifically XAI can facilitate an alleged discovery. This uncertainty directs to an area of potential elaboration about what qualifies as XAI for

discovery, and which developments have been made in that regard.

Hence, this paper aims to synthesise existing literature centring around recent developments of XAI. However, rather than exploring all facets of XAI in the current landscape, there will be emphasis on use of XAI for the purpose of discovery.

To do so, the following study will methodically collect relevant documents regarding XAI for discovery, investigating the different XAI methods employed and the ways in which the notion of 'discovery' is understood across the retrieved literature.

2 Methods

The literature search was conducted using a focused search string to source the relevant documents from the Scopus scientific database. Based on exploratory investigation on Google Scholar regarding 'Explainable AI for Discovery' and referenced supplementary literature, the following search query was formed:

(Interpretab* OR Explainab*) AND ("Artificial Intelligence" OR AI) AND "Knowledge Discovery"

'Interpretab*' OR 'Explainab*' have been selected as representative keywords with wildcard expansions through the use of asterisks. Shortening the terms to 'Interpret*' OR 'Explain*' was flagged as a liability, since they are both common verbs outside of their keyword contexts, which increased the search scope to include irrelevant documents. The chosen varieties allow for flexibility between, for example 'Explainable' and 'Explainability', both of which are fully compliant with the focus of this review. 'Interpretab*' was also included as an alternate keyword due to its interchangeability with 'Explainab*', particularly within ML spaces. While certain distinctions have been discussed for the two terms, overlap between their definitions remains, necessitating consideration of both options (Barredo et al., 2020).

"Artificial Intelligence" OR AI, as the overall domain of research, has been included in both its full-text and initialism formats due to increased integration of the shortened version in common vocabulary. Therefore, in case researchers forego the full words, their work can still be included in the search.

As the primary interest, 'Knowledge Discovery' as a whole has been included for further domain specificity. 'Discover*', as a singular word, is not an established domain of research and is instead a common descriptor, which opens the scope of the search to tangential or irrelevant papers, as tested in previous search string iterations. 'Knowledge Discovery', however, is an established aim for pattern discovery, with data mining at its core, but later developed more diversified applications as technology and data complexity increased (Shu & Ye, 2023). Therefore, its utilisation maintains the focus of the search without being overly involved.

The identification, screening, and selection process for the final search string results were documented in accordance with the PRISMA standard (PRISMA, 2024), as presented in Figure 2.1. In the identification process, 1 duplicate document was removed and automation tools, namely the native Scopus filters, were employed to remove an additional 72. The filters applied consisted of year range, document type, and language. While the non-filtered scope of the search ranged from 2005 to 2026, field-wide understanding of what XAI entails has developed and diverged over time. Therefore, the search was confined between 2019 and 2026 (Figure 2.2). These dates encompass the period of rapid increase in relevant papers, and allow for exploration of literature post-categorisation of XAI purposes by Adadi & Berrada (2018). Furthermore, the search was restricted to 'Article' and 'Conference Paper' document types, notably removing reviews to lessen recursive meta-analyses and retain the focus of the current review. Lastly, the language filter was set to English to ensure fair interpretation of the documents without third-party translation.

After the initial identification process, 125 records remained for manual screening. In this screening, the abstract and metadata of the documents were assessed. Three main exclusion criteria were employed to determine further removals. The *first* was whether or not the document is a review. As aforementioned, reviews were automatically filtered out of the search; however, some were not tagged as such and required manual removal. The *second* criterion was explicit use of explainable AI. While the search string accounted for alternative key-term usage, there are cases of discussing interpretable or explainable AI without it being the

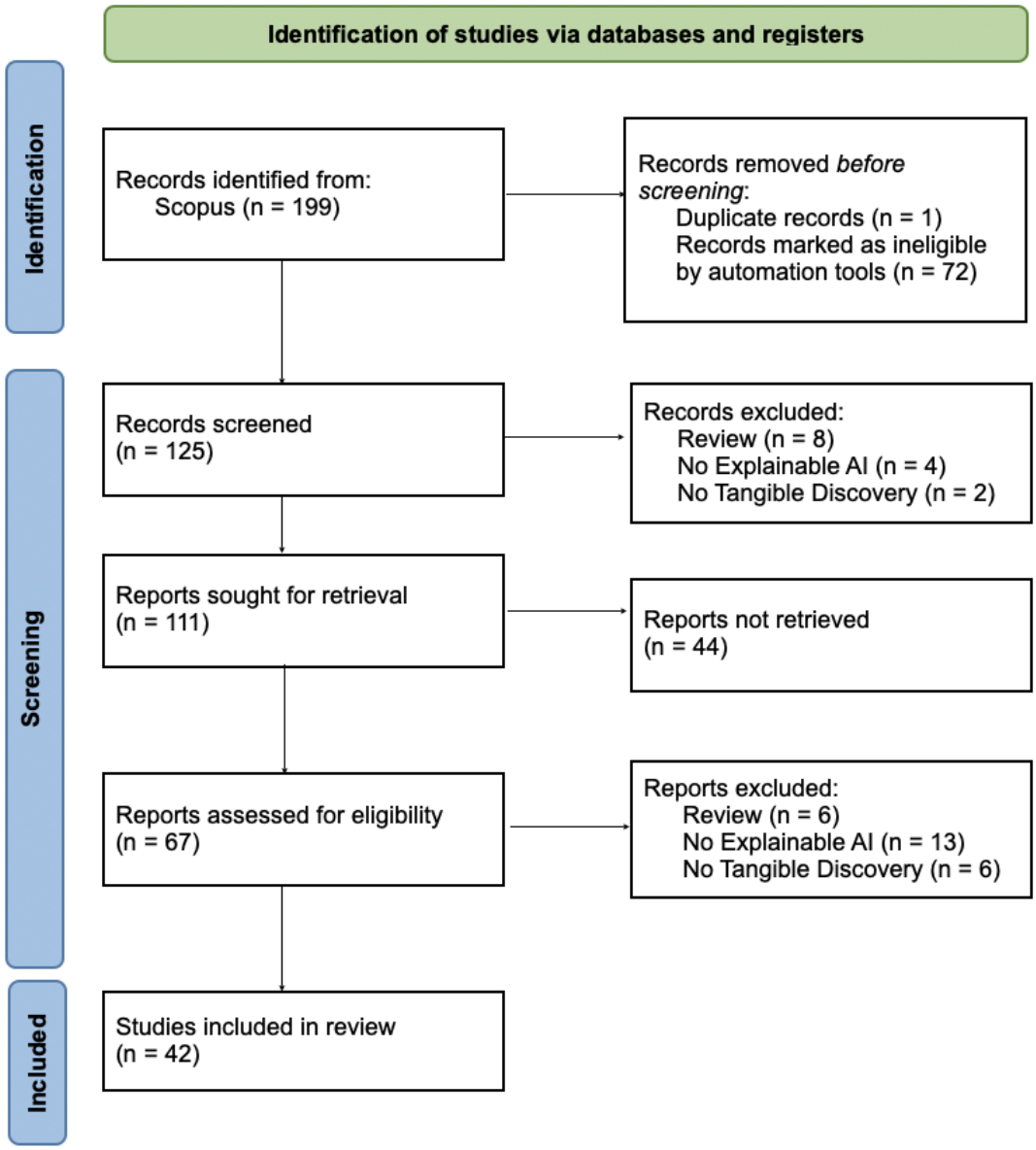

Figure 2.1: In the selection of a document seed set, the PRISMA standard for systematic review was followed. Each decision made at each step of identification, screening, and inclusion is documented in the flowchart as evidence of a structured selection process and to support reproducibility of the methods.

main point of research. Therefore, it was necessary to ensure that some method of XAI is the primary focus of each paper. The *third* criterion concerned discovery. Papers often include potential for discovery within their conclusion or future work section. However, it is variable whether there is a clear through-line towards a discovery or tangible understanding of what that discovery may entail. Thus, this criterion ensures that discovery is not briefly mentioned as an afterthought, but instead is integral to the story of the paper. From the remaining 111 documents, 44 were then removed due to lack

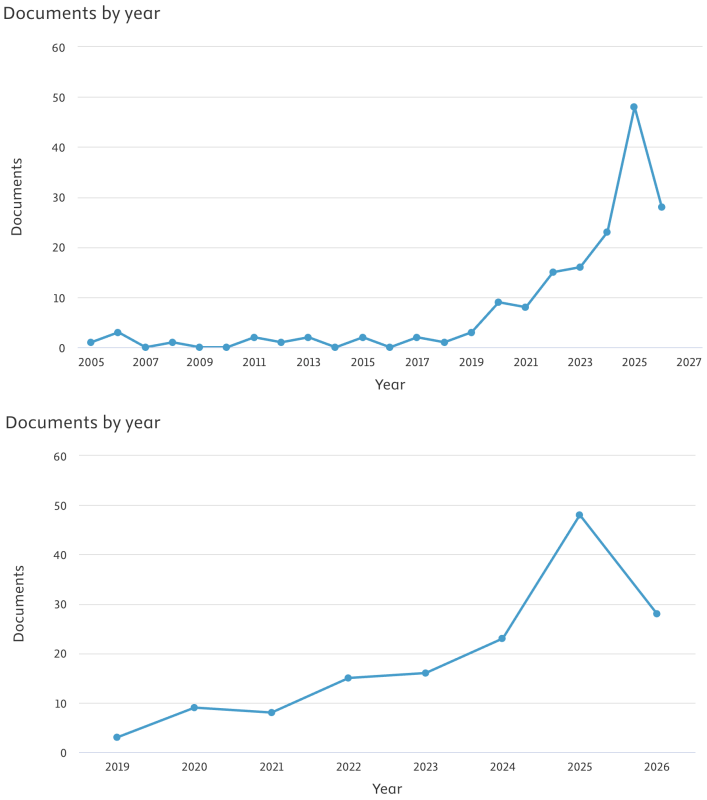

Figure 2.2: The upper graph reports the publishing years for all documents retrieved by the search query, while the the lower graph zooms in on the chosen range for the search, 2019-2026, for additional clarity. Both graphs illustrate an upward trend in relevant papers within the chosen range, reflecting the recent surge in XAI research for discovery.

of institutional access.

The final screening step consisted of assessing the full documents for the aforementioned exclusion criteria. Primarily, the introduction, discussion, and conclusion sections were referenced, with further examination of the methods section if additional clarity was needed. This process concluded with 42 accepted papers (Appendix A).

3 Results

This investigation synthesised two core areas across the selected documents: XAI and knowledge discovery. In the area of XAI, I examine the different methods utilised in each paper with the aim to make a discovery, organising them based on type of approach and technologies used. Meanwhile, the area of knowledge discovery pertains to the findings themselves, wherein I explore the different ways in which discovery is attributed throughout the sources.

3.1 XAI Methods

The assessed literature employs a variety of different XAI methods to facilitate discovery (Table 3.1). Approaches to implementing XAI can be categorised into two primary subsections: interpretable and post-hoc methods. Regarding interpretable models, the models themselves can be understood based on factors such as algorithmic transparency, decomposability, and simulatability (Barredo et al., 2020). Contrastingly, post-hoc methods are used for models that are not inherently explainable. Therefore, techniques are applied to the model after its

deployment in an effort to inform about the model's findings or conclusions.

3.1.1 Methods Towards Model Interpretability

Tree-based machine learning methods utilise boolean-like logic to predict the value of a target variable against input variables. The structured approach allows for inherent interpretability, often used to inform decision-making processes (Dwivedi et al., 2023). Further modifications to the model's use can be applied for increased transparency. One such modification can be made during a key machine-learning step, feature selection. Lötsch & Ultsch (2022) presented an Explainable Feature Selection (XFS) approach, wherein previously unselected features were methodically revisited with the aim to balance classification accuracy with informative features. Another measure is to implement an explainability factor into the model itself. Bobek & Nalepa (2021) employs uncertain decision trees to provide insight into reasoning behind the decisions made.

Furthermore, Waghen & Ouali (2019) proposed Interpretable Logic Tree Analysis (ILTA), which combines knowledge discovery in databases with fault tree analysis to create an interpretable logic tree. However, when multiple trees are utilised in an ensemble, the model becomes increasingly opaque, necessitating additional explainability measures (Dwivedi et al., 2023). An example of this is the use of gradient-boosted trees, wherein residual errors are iteratively corrected (Gerevini et al., 2023; Kırboğa et al., 2023; Zhe et al., 2025; L. Xu et al., 2025). In this case, and for other tree-based cases which require additional insight to be explainable (Gomez-Cravioto et al., 2022; Gabardo et al., 2025; Dagnino et al., 2026; Gherardini et al., 2024), further post-hoc analysis (Section 3.1.2) or complex model combinations (Section 3.1.3) are required.

Knowledge Graphs (KGs) provide a network of facts and relationships, which can be linked to ontologies of background information to derive new findings (Dwivedi et al., 2023). They can be combined with rule-based or ML methods to facilitate explainability. Some examples include the integration with aforementioned decision-tree models and Neural Networks (NNs) (Talukder et al., 2020). The subsymbolic nature of NNs is a key contribution

Table 3.1: Distribution of Methods Across Sources.

Method	Source
Tree	Lötsch & Ultsch (2022); Bobek & Nalepa (2021); Waghen & Ouali (2019); Gerevini et al. (2023); Kırboğa et al. (2023); Zhe et al. (2025); L. Xu et al. (2025); Gomez-Cravioto et al. (2022); Gabardo et al. (2025); Dagnino et al. (2026); Gherardini et al. (2024)
Graph	Talukder et al. (2020); Kalutharage et al. (2024); Maree (2025); Sharma & Jain (2024); Saluja & Bhavsar (2025); Papageorgiou et al. (2025); Gherardini et al. (2024)
Other (Interpretable)	Shahrzad et al. (2022); Almeida et al. (2023); H. Xu et al. (2025); Heinrich et al. (2025); De Santis et al. (2021); D. Wang et al. (2024); Y. Xu et al. (2025); Huang et al. (2026)
Deep Learning	Iwasaki et al. (2022); Thunold et al. (2023); Kim et al. (2025); Han & Chang (2023); Klein et al. (2024); Lu et al. (2024); Withnell et al. (2021); Huang et al. (2026); Kalutharage et al. (2024); Tarnowska et al. (2022); Mühlberg et al. (2023); Wagner et al. (2024); Ehrensperger et al. (2025); Gerevini et al. (2023); Gabardo et al. (2025)
SHAP	Bobek & Nalepa (2021); Gerevini et al. (2023); Kırboğa et al. (2023); Zhe et al. (2025); Y. Xu et al. (2025); Gomez-Cravioto et al. (2022); Gabardo et al. (2025); Dagnino et al. (2026); Kalutharage et al. (2024); Han & Chang (2023); Withnell et al. (2021); Mühlberg et al. (2023); Wagner et al. (2024); Ehrensperger et al. (2025); X. Wang et al. (2022); Vlahek & Mongus (2021); H. Xu et al. (2025)
Other (Post-Hoc)	Kim et al. (2025); Gâlmeanu et al. (2025); Gâlmeanu & Andonie (2025); Naser (2022); Wagner et al. (2024); Gherardini et al. (2024); Angerri et al. (2025)

to their black-box reputation. Hence, through integration with a KG, the models can be informed by meaningful symbolic representations, leading to informed insights into the previously opaque NN by creating a Neuro-Symbolic (NS) system (Dwivedi et al., 2023). Examples of this include works by Kalutharage et al. (2024) and Maree (2025), who developed NS models to detect cyberattacks and explore cultural heritage data respectively.

Furthermore, KGs can be linked to Large Language Models (LLMs) to be queried and provide explanations in a manner which is understandable to non-field experts. Sharma & Jain (2024) presented OntoXAI, which combines Semantic Web Rule Language (SWRL) and a KG to explain dengue disease classification. Maree (2025) leverages LLMs to include contextual awareness in explanations of graph-relations, based on an interestingness measure. Saluja & Bhavsar (2025) created a Knowledge Discovery Framework for Cloud Forensic Investigation (KDCFI), harnessing LLMs to query cloud forensic evidence. Papageorgiou et al. (2025) developed a graph-based version of Retrieval Augemented Generation (RAG), which supplements an LLM with key domain-specific documents. Lastly, Gherardini et al. (2024) incorporated KGs into a combined method (Section 3.1.3).

Other pursuits towards interpretable models include novel methods like Evolution of Transparent Explainable Rule-sets (EVOTER) (Shahrzad et al., 2022), an Explainable Boosting Machine (Almeida et al., 2023), Genetic Algorithm-Geographic Weighted Regression (GA-GWR) (H. Xu et al., 2025), and a Audio Prototypical Part Network (AudioProtoPNet)(Heinrich et al., 2025). Furthermore, De Santis et al. (2021) implemented granular computing with text categorisation and support vector machines, D. Wang et al. (2024) implemented genetic symbolic regression (SR), Y. Xu et al. (2025) implemented an automated theorem generator, and Huang et al. (2026) implemented a Denoising Convolutional Autoencoder (DCAE) (Section 3.1.3).

3.1.2 Post-Hoc Methods

While Deep Learning (DL) algorithms are not inherently considered transparent, post-hoc methods can be applied to facilitate interpretability (Dwivedi et al., 2023). When discussing DL, Deep Neural Networks (DNNs) are commonly the model of focus. DNNs utilise multiple hidden layers to learn complex patterns, making their inner-workings challenging to explain. One approach is to create visual explanations of the DNN's findings, as implemented by Iwasaki et al. (2022); Thunold et al. (2023); Kim et al. (2025); Han & Chang (2023); Klein et al. (2024); Lu et al. (2024). Respectively, Iwasaki et al. (2022), Thunold et al. (2023), and Kim et al. (2025) revealed clusters to explore SARS-CoV-2 genome composition, classify polyps, and support precipitation forecasts, the latter two layered over image data. Han & Chang (2023) combined reconstruction error distribution and t-distributed stochastic neighbour embedding (t-SNE) to explain the reconstruction error in an adaptive semi-supervised DNN. Klein et al. (2024) utilised counterfactual explanations (CEs) and the Testing of Concept Activation Vectors (TCAV) to test various input observations and test concept importance to each DNN layer. Lu et al. (2024) created a novel isolation-integration strategy for feature visualisation to reveal feature heatmaps over electrocardiogram (ECG) data.

Autoencoders are a type of DNN which learn in an unsupervised manner, implemented by Withnell et al. (2021); Huang et al. (2026); Kalutharage et al. (2024). Where Withnell et al. (2021) utilised post-hoc explanations and Huang et al. (2026) created a model interpretability approach. Kalutharage et al. (2024), as well as Tarnowska et al. (2022), created more complex model combinations (Section 3.1.3). DNNs can otherwise be enhanced through post-hoc feature analysis, such as in Self-Enhancing Multi-Photon Artificial Intelligence (SEMPAI) (Mühlberg et al., 2023), ECG analysis (Wagner et al., 2024), analysing ERA5 soundings (Ehrensperger et al., 2025), prognosis estimation for COVID-19 (Gerevini et al., 2023), and fatal traffic outcome investigation (Gabardo et al., 2025); the following paragraph providing further feature-based XAI elaboration.

The most prevalent post-hoc method in the reviewed literature is SHapley Additive exPlanations (SHAP). SHAP encompasses the features that lead to a model's output, providing information about how a prediction has been found and the extent to which different features contribute to that prediction. SHAP values can be calculated alongside a variety of models, with the ability to provide insight

into global and local feature importance (Dwivedi et al., 2023). From the aforementioned models, (gradient boosted) trees (Bobek & Nalepa, 2021; Gerevini et al., 2023; Kırboğa et al., 2023; Zhe et al., 2025; Y. Xu et al., 2025; Gomez-Cravioto et al., 2022; Gabardo et al., 2025; Dagnino et al., 2026), graphs (Kalutharage et al., 2024), and DL models (Han & Chang, 2023; Withnell et al., 2021; Kalutharage et al., 2024; Mühlberg et al., 2023; Wagner et al., 2024; Ehrensperger et al., 2025; Gerevini et al., 2023) are commonly integrated with SHAP. Otherwise, SHAP can also be combined with methods like Seq2Pat, a sequential pattern mining tool (X. Wang et al., 2022), iterative feature generation (Vlahek & Mongus, 2021), and GA-GWR (H. Xu et al., 2025).

Other post-hoc methods include focus on visualisation, (Kim et al., 2025; Gâlmeanu et al., 2025; Gâlmeanu & Andonie, 2025), SHAP adjacent methods (Naser, 2022; Wagner et al., 2024; Gherardini et al., 2024), and links from Principle Component Analysis (PCA) to metadata (Angerri et al., 2025).

3.1.3 Combined Methods

More complex method combinations can also be developed to synthesise explainability from a multitude of sources. In enhancing tree-based methods, Bobek & Nalepa (2021) created Local Uncertain Explanations (LUX) to measure certainty of explanations, utilising decision trees and various post-hoc methods including SHAP; meanwhile Dagnino et al. (2026) combined PCA, random forest, and SHAP to analyse patient complexity profiles to examine contributing factors of depression.

When combining with KGs, Kalutharage et al. (2024) combined autoencoders, KGs, and SHAP to detect Internet of Things attacks, and Gherardini et al. (2024) created a Comprehensive Abstraction and Classification Tool for Uncovering Structures (CACTUS), which combines decision trees, KGs, and post-hoc feature plotting.

In combination with DL methods, Tarnowska et al. (2022) utilised a knowledge-based framework for clinical decision support system (CDSS), a Supervised NN, association and action rules, and a knowledge encoder for tinnitus retraining therapy, and Huang et al. (2026) developed a Denoising Convolutional Autoencoder (DCAE) using DNNs and RAG.

3.2 Level of Discovery

Among the assessed papers, various levels of depth are attributed to discovery within the context of each research focus. The depth varies between low, medium, and high levels of discovery (Table 3.2), based on their respectively defined criteria.

Table 3.2: Distribution of Sources Across Discovery Levels.

Level	Source	Count
Low	Bobek & Nalepa (2021); Shahrzad et al. (2022); Lötsch & Ultsch (2022); De Santis et al. (2021); Vlahek & Mongus (2021); Maree (2025); Y. Xu et al. (2025); Angerri et al. (2025)	8
Medium	Waghen & Ouali (2019); Talukder et al. (2020); Withnell et al. (2021); X. Wang et al. (2022); Iwasaki et al. (2022); Naser (2022); Gomez-Cravioto et al. (2022); Tarnowska et al. (2022); Gerevini et al. (2023); Mühlberg et al. (2023); Kırboğa et al. (2023); Thunold et al. (2023); Han & Chang (2023); Kalutharage et al. (2024); Sharma & Jain (2024); Wagner et al. (2024); Gherardini et al. (2024); Klein et al. (2024); Lu et al. (2024); Heinrich et al. (2025); Saluja & Bhavsar (2025); Zhe et al. (2025); Kim et al. (2025); L. Xu et al. (2025); Ehrensperger et al. (2025); Gâlmeanu et al. (2025); Papageorgiou et al. (2025); Gâlmeanu & Andonie (2025); Dagnino et al. (2026); Gabardo et al. (2025); Huang et al. (2026)	31
High	Almeida et al. (2023); D. Wang et al. (2024); H. Xu et al. (2025)	3

3.2.1 Low Level

The paper presents a tool that can facilitate research at one of the higher levels of discovery, such as by modifying a component of the XAI method itself for ease of interpretability or improved explanations.

At the low level, the findings are not directly novel, but can be utilised as an underlying factor towards a discovery by making methods more robust or relationships more apparent.

By developing LUX, Bobek & Nalepa (2021) applied uncertainty into XAI, which can facilitate discovery through previously unknown insight into the model's decision process.

The aim of EVOTER was to create explainable rule sets to uncover relationships for a variety of use cases, such as blood pressure prediction. These rules can be used to discover which features affect others and the manner in which that effect occurs; this could discover new information about the given task (Shahrzad et al., 2022).

Three papers focused on feature selection, namely XFS (Lötsch & Ultsch, 2022), iterative feature learning (Vlahek & Mongus, 2021), and information granulation approach to text classification (De Santis et al., 2021). In XFS, features that may have been overlooked could be reconsidered for the machine learning process. In iterative feature learning, the new features were generated and statistically assessed to agree with existing features. In the information granulation approach, a genetic algorithm was employed to select relevant m-grams for the classification task. These methods can be employed as a component of a larger model to uncover findings that otherwise would have remained hidden due to the lack of meaningful feature integration.

Y. Xu et al. (2025) created a method called '$\Delta 1$', which generated logically sound, mutually non-equivalent theorems based on input propositional literals or first order logic predicates. In this, new logical relationships can be revealed, which can be utilised as underlying step for further discovery.

With a focus on uncovering hidden connections, Maree (2025) allowed for the traversal of symbolic knowledge graphs through measures of interestingness, and Angerri et al. (2025) leveraged PCA in combination with metadata to link variables across conceptual domains. In exploring these relationships, interpretation can be used to formulate a line of reasoning towards a discovery.

3.2.2 Medium Level

The paper presents tangible potential for discovery. This potential is determined to be tangible based on findings that align with existing knowledge and a clear path towards what could be discovered through its use.

At the medium level, while a novel discovery may not have been made within the scope of the paper, the XAI methods are supported by empirical findings. Thus, providing a basis to predict that further application of the proposed methods could bring new knowledge to light.

Talukder et al. (2020) and Tarnowska et al. (2022) presented decision support systems for healthcare, focused on balancing reactive and proactive care, and tinnitus support therapy respectively. Additionally, Dagnino et al. (2026) built patient complexity profiles for depression, analysing the contributing features for different severity levels. With theses systems, doctors can obtain diagnostic insights that can influence the manner of treatment provided.

A prevalent research topic is the exploration of genetic mutations, health conditions, or viruses, such as Withnell et al. (2021) who utilised XOmiVAE to reveal the contribution of each gene towards tumour prediction or Wagner et al. (2024) who implemented an unsupervised learning system to identify heart-attack subgroups.

In particular, several papers presented Covid-19-centric systems, namely Iwasaki et al. (2022) who implemented BLSOM to reveal mutations contributing to Covid-19, Gerevini et al. (2023) who explored features contributing to Covid-19 prognosis prediction, and Kırboğa et al. (2023) who used CardioVascular Diseases 22 (CVD22) to understand influences on Troponin levels during infection.

In further medical application, Sharma & Jain (2024) developed OntoXAI, which combined a knowledge base of symptoms, SWRL rules, and an LLM interface to classify dengue fever with natural language explanations and post-hoc feature rankings. Lu et al. (2024) combined DNNs and visualisations to interpret ECGs. All of these health-related methods have the potential reveal features that contribute to detrimental illnesses, discovering new paths towards more effective care.

Both Mühlberg et al. (2023) and Thunold et al. (2023) developed systems for imaging-based biomarker discovery. In doing so, they were able to uncover factors contributing to damage in muscle fibre and gastrointestinal polyps from visual data, which could lead to deepened diagnostic understanding.

Waghen & Ouali (2019) and Naser (2022) developed systems for causality analysis through ILTA to reveal faults in a system and mapping functions to derive causal expressions for structural engineering, respectively. In doing so, the uncovered causal knowledge can be used to predict events and structural capacities.

Several papers examined features of societal phenomena through observed patterns. X. Wang et al. (2022) implemented Seq2Pat for customer intent prediction, Gomez-Cravioto et al. (2022) highlighted variables that affect alumni income, and Gabardo et al. (2025) investigated features that contribute to fatalities in car crashes. By revealing contributing features to these fields, they could discover which factors can be leveraged for desired outcomes, such as consumer targeting, pursuit of higher income, and reduced fatal crashes.

In the realm of cybersecurity, Kalutharage et al. (2024), Huang et al. (2026), Han & Chang (2023), and Saluja & Bhavsar (2025) focused on intrusion detection, aiming to detect, interpret, and respond to attacks. In doing so, they can learn previously unknown attack patterns and develop responses to ensure maintained security.

In the field of natural sciences, Kim et al. (2025) and Ehrensperger et al. (2025) researched concept discovery for extreme weather and lightning processes respectively. In doing so, they were able to distinguish between different visually similar weather mechanisms and identify atmospheric processes that contribute to lightning activity. These insights can be used to discover new contributions to different extreme weather conditions.

L. Xu et al. (2025) and Zhe et al. (2025) brought focus to geology. The former enabled the revelation of contributing features to permeability in complex tight sandstone structures, while the latter predicted grout volume with feature analysis to discover interactions between geological features and other contributing factors.

Other natural scientific studies include Klein et al. (2024) and Heinrich et al. (2025). Klein et al. (2024) used TCAV and CEs to understand the relationship between perovskite thin-film formation and solar cell performance, the insights of which can used to further the manufacturing of solar cells. Heinrich et al. (2025) classified bird calls using AudioProtoPNet, providing global prototype explanations and local similarity explanations, enabling insight into bioacoustic data.

With focus on the visualisation of concept drift, Gâlmeanu et al. (2025) and Gâlmeanu & Andonie (2025) first explored how features affect concept drift, then how to detect concept drift. These investigations provide insight into multiple facets of concept drift, with supporting visualisations aiding the discovery of data dynamics.

Further miscellaneous domains include government, and data abstraction and classifications. Papageorgiou et al. (2025) created a GraphRAG system for E-Government. In this, they can utilise question-answering to discover synthesised information regarding government data. Gherardini et al. (2024) developed CACTUS, a tool to support classification of categorical attributes in smaller datasets through knowledge graphs and feature ranking, providing insight into classes and interactions for better data analysis.

3.2.3 High Level

The paper presents an explicit discovery made through the use of XAI that demonstrates the research's novelty.

At the high level, the researchers claim to have made a novel discovery that is facilitated by XAI use.

Almeida et al. (2023) developed a decision support system for treatment of Acute Myeloid

Leukemia (AML). They utilised an SVM for feature selection and EBM, an interpretable model relying on rules for increased transparency, in combination with an ensemble, to predict a patient's survival. They created models for each dataset: clinical, gene expression, and gene mutation, as well as all combinations. In doing so, they found contributing clinical features and gene mutation features supported by existing literature, such as the patient's age or mutations in the TP53 gene. However, they also found mutations in the PHF6 gene and genetic expression features of KIAA0141, MICALL2, and SLC9A2 to be highly influential in the prediction of AML, with no existing links in supporting AML literature. Notably, they have been linked to other cancer-related diagnoses; therefore, a newfound link to AML could lead to new developments towards diagnosing and treating the illness.

Through SR, D. Wang et al. (2024) discovered an interpretable expression for the mean velocity deficit throughout the wake of a wind turbine. Existing wake models stem from simplified expressions, unable to capture all nuances of turbine wakes. In contrast, the SR expression with the greatest determined fitness aligns with ground truth data, able to capture the full wake and its velocity deficit over time. In doing so, the expression represents how the velocity deficit fluctuates along the stream-wise direction, the fixed nature of peaks in a span-wise direction, and the fast expansion of the wake in the downstream far-wake region under low turbulence. In the accurate representation of observed wake data, these findings support the capacity to overcome idealised assumptions in existing theoretical frameworks through XAI-backed discovery, facilitating increased understanding of physical mechanisms.

H. Xu et al. (2025) utilised GA-GWR to derive equations that describe the relationship between precipitation patterns and terrain characteristics. To do so, they combined Gaussian weighted regression with genetic algorithms, allowing for spatial analysis and discovery of the fittest formula as represented by genomes. The resulting spatio-temporal formulas were used to analyse the impact of climate change on various regions, providing insight into the extent to which terrain-precipitation relationships have been affected. Furthermore, they were able to embed their equations in predictive models to predict long-term climate effects on precipitation, allowing for the development of projections that can aid in disaster prevention.

4 Discussion

4.1 Statement of Positionality

Before beginning the discussion, it is important to note my stance as an author and the motivation behind my perspective. As a Bachelor's Student studying Artificial Intelligence, I have background in the AI-specific aspects of the retrieved literature. However, there are areas which I am interpreting as an outsider, particularly specific applications in Medicine, the Natural Sciences, Engineering, and Applied Mathematics; hence, I may be unaware of certain nuances that domain experts would find notable.

4.2 The Role of XAI in Discovery

By investigating the variety of XAI methods and the resulting discoveries, we can contemplate the extent to which the interpretable or explainable portion of the model are responsible for the discovery made.

As the most prominently implemented XAI method (Table 3.1), SHAP demonstrates how an explainability tool directly produces the alleged discovery. In this, the papers where SHAP is the primary XAI method present feature contributions as the discovery itself. Notably, all but one of these papers are congregated within the *Medium Level of Discovery* (Section 3.2.2), as the feature-based explanations provided by SHAP, while insightful and well-supported, did not amount to explicitly novel findings. The one outlier was placed within the *High Level of Discovery* (Section 3.2.3), but SHAP was not the sole XAI method implemented (H. Xu et al., 2025). However, without the usage of SHAP, the models' predictions would be opaque, lacking comprehension of how the conclusions were formed. This would prevent any level of discovery, unless the prediction itself is new knowledge; but without model transparency the legitimacy of that knowledge would continue to be questioned. Hence, XAI plays a key role in discovery, as a tool better understood can be better leveraged, whether that be within an inherently interpretable model or a

post-hoc explanation such as SHAP.

4.3 Reliability of Methods

In advocating for XAI methods to be utilised for the purpose of discovery, it is important to acknowledge the reliability of the methods themselves. As seen in Table 3.1, there is a spread of both interpretable and post-hoc methods used, with some overlapping combinations.

Post-Hoc methods in particular are subject to criticism due to their strong assumptions and misleading confidence (Bobek & Nalepa, 2021). It can be argued that they are not truly faithful to the model's workings due to only being approximations after predictions have already been formed (Heinrich et al., 2025; Klein et al., 2024).

Inherently interpretable methods are regarded as more faithful. However, while not always the case, they are often associated with weaker performance accuracy than their black box counterparts (Vlahek & Mongus, 2021), with difficulty adapting and handling complexity (Kalutharage et al., 2024; Maree, 2025).

If an XAI method is not representative of the model's true workings, then the validity of presented discoveries should be put under scrutiny. While the studies of focus are critical of baseline post-hoc and interpretable methods as a means to argue for their research gap, the final methods and findings are not afforded the same level of balanced analysis. Of the assessed papers, only 11 explicitly acknowledge the limitations of their own presented methods (Gomez-Cravioto et al., 2022; Tarnowska et al., 2022; Lötsch & Ultsch, 2022; Mühlberg et al., 2023; Kırboğa et al., 2023; Wagner et al., 2024; Klein et al., 2024; Saluja & Bhavsar, 2025; Papageorgiou et al., 2025; Dagnino et al., 2026; Huang et al., 2026). This indicates a greater problem of potentially misrepresenting findings as more novel or prone to discovery than they realistically are. While this research issue is not necessarily unique to the field of XAI, its prevalence in XAI-specific papers does highlight the idealisation of new technologies, the progress of which should instead be assessed from a measured perspective.

4.4 Discovery as a Concept

An area of contention is the understanding of discovery itself. While the proposed taxonomy organises the discoveries into hierarchical levels based on tangibility, there is value in observing the manner in which discoveries are presented across the gathered literature.

The first observation is solely utilising 'discover*' as a keyword in the abstract of the paper without any further mention throughout the body of work. In doing so, discovery is leveraged as a tool for search optics, rather than a research focus or outcome.

The next is 'discover*' as introductory background. In this case, the presentation of discovery is often associated with established research fields such as Knowledge Discovery in Databases (KDD) (Shu & Ye, 2023), demonstrating a broad overview of the field-wide research aim, but not necessarily the role of discovery in the conducted study.

Another observation is the methods-focused use of 'discover*', which refers to technical components. In this, the papers aim to present which facet of their methodology is facilitating the discovery, often interweaving with the XAI implementation.

Lastly, incorporating the discovery in the later sections of the paper, such as in the discussion and conclusion, has two primary approaches. One approach is to present what was discovered specifically, examining which findings are novel and indicative of successful discovery. The other approach is to allude to future discovery potential, expressing that a discovery could be made on the basis of the methods and findings.

The inconsistency with which discovery is presented brings into question the importance of a discovery consensus in the field. On one hand, without an agreement of what a discovery truly is, the importance may become diluted due to ever-increasing claims to have made a discovery using XAI, taking advantage of the current ambiguity. This in combination with the struggle for scientific standards to keep up with rapid technological development may lead to irresponsible scientific reporting and practices. However, to confirm an agreed upon definition of discovery, it would be critical to determine who verifies that a discovery can indeed be categorised as such, and whether having a verification process places unnecessary re-

striction on research endeavours, counteracting the advocation for more open and accessible scientific practices.

Hence, the presented taxonomy for organising by *Level of Discovery* (Section 3.2) exemplifies a potential avenue for developing understanding of discovery on a spectrum rather than a concrete definition. Considering the multitude of perspectives projected onto discovery within the reviewed set of documents, developing a singular definition which encompasses all possibilities would be greatly challenging. As aforementioned, requiring an authoritative body to verify discovery may be restrictive, with complications in determining who should govern that decision due to the innate interdisciplinary nature of discovery. Rather, providing a spectrum of discovery for reference can aid in documenting which perspective on discovery is the focus of a given work, allowing for clarity while also preserving diversity in research endeavours.

4.5 Limitations and Future Work

As stated in Section 4.1, as the sole curator of this review, I view the literature through an AI lens. While this is fully applicable to XAI technology, certain nuances within specific domains and the understanding of knowledge discovery may be lacking. Furthermore, conducting a review on my own restricted the breadth of sources that could be reasonably assess and the depth which could be dedicated to each of them. Therefore, as an extension to this review, a group of contributors from diverse backgrounds could work together when building upon the proposed taxonomy of discovery. In doing so, an interdisciplinary perspective can be consulted, such as insight from a philosophical lens on what constitutes as new knowledge, to extend the proposed discovery spectrum with additional works and characteristics per discovery category.

5 Conclusion

Overall, XAI for knowledge discovery is a field in which interest is continuing to grow, with more complex models necessitating more complex methods to be understood. By examining the patterns within existing research, we can gain an overview of the most prevalent XAI methods and their role in the multitude of discoveries presented. However, result embellishment and lack of consensus runs the risk of overestimating our achievements without grounded verification. Hence, moving forward we should bring a critical perspective to the proposed methods and a diverse outlook on discovery as a spectrum of findings, ensuring that we make responsible steps towards scientific advancements through XAI-backed discovery.

References

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE access*, *6*, 52138–52160.

Almeida, J. M., Castro, G. A., Machado-Neto, J. A., & Almeida, T. A. (2023). An explainable model to support the decision about the therapy protocol for AML. In *Brazilian conference on intelligent systems* (pp. 431–446).

Angerri, X., Sacco, D., Bosque, E., & Gibert, K. (2025). Automatic labelling of topics in topic modelling. In *Artificial intelligence research and development* (pp. 69–78). IOS Press.

Barredo, A. A., Del Ser, J., Gil-Lopez, S., Díaz-Rodríguez, N., Bennetot, A., Chatila, R., ... others (2020). Explainable Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information fusion*, *58*, 82–115.

Bobek, S., & Nalepa, G. J. (2021). Introducing uncertainty into explainable AI methods. In *International conference on computational science* (pp. 444–457).

Dagnino, P., Salinas, M., & Salas, R. (2026). Patient complexity profiles in depression: a machine learning approach to personalized mental health. *Frontiers in Psychiatry*, *17*, 1741860.

De Santis, E., Capillo, A., Ferrandino, E., Mascioli, F. M. F., & Rizzi, A. (2021). An information granulation approach through m-grams for text classification. In *International joint conference on computational intelligence* (pp. 73–89).

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., ... others (2023). Explainable AI (XAI): Core ideas, techniques, and solutions. *ACM computing surveys*, *55*(9), 1–33.

Ehrensperger, G., Simon, T., Mayr, G. J., & Hell, T. (2025). Identifying lightning processes in era5 soundings with deep learning. *Geoscientific Model Development*, *18*(4), 1141–1153.

Gabardo, A. M., Schünemann, G. A., Jaskowiak, P. A., Moreira, B. G., & Pfitscher, R. J. (2025). Highway to... determining fatal outcomes in traffic accidents based on police reports. In *Brazilian conference on intelligent systems* (pp. 230–244).

Gâlmeanu, H., & Andonie, R. (2025). Concept drift detection and visualization with shifting window. *Information Visualization*, *24*(4), 412–428.

Gâlmeanu, H., Kovalerchuk, B., & Andonie, R. (2025). Interactive discovery of concept drift with lossless visualization in machine learning. In *International conference on human-computer interaction* (pp. 310–324).

Gerevini, A. E., Maroldi, R., Olivato, M., Putelli, L., & Serina, I. (2023). Machine learning techniques for prognosis estimation and knowledge discovery from lab test results with application to the COVID-19 emergency. *IEEE Access*, *11*, 83905–83933.

Gherardini, L., Varma, V. R., Capała, K., Woods, R., & Sousa, J. (2024). CACTUS: a comprehensive abstraction and classification tool for uncovering structures. *ACM Transactions on Intelligent Systems and Technology*, *15*(3), 1–23.

Gomez-Cravioto, D. A., Diaz-Ramos, R. E., Hernandez-Gress, N., Preciado, J. L., & Ceballos, H. G. (2022). Supervised machine learning predictive analytics for alumni income. *Journal of Big Data*, *9*(1), 11.

Han, Y., & Chang, H. (2023). XA-GANomaly: An explainable adaptive semi-supervised learning method for intrusion detection using GANomaly. *Computers, Materials and Continua*, *76*(1), 221–237.

Heinrich, R., Rauch, L., Sick, B., & Scholz, C. (2025). AudioProtoPNet: An interpretable deep learning model for bird sound classification. *Ecological Informatics*, *87*, 103081.

Huang, Y., Ma, M., Raymond, W. J. K., & Chow, C.-O. (2026). An explainable and adaptive internet of things intrusion detection system supported by large language models. *Engineering Applications of Artificial Intelligence*, *163*, 112911.

Iwasaki, Y., Abe, T., Wada, K., Wada, Y., & Ikemura, T. (2022). Unsupervised explainable AI for molecular evolutionary study of forty thousand SARS-CoV-2 genomes. *BMC microbiology*, *22*(1), 73.

Kalutharage, C. S., Liu, X., Chrysoulas, C., & Bamgboye, O. (2024). Neurosymbolic learning in the XAI framework for enhanced cyberattack detection with expert knowledge integration. In *IFIP international conference on ict systems security and privacy protection* (pp. 236–249).

Kim, S., Choi, J., Lee, S., & Choi, J. (2025). Unsupervised concept discovery for deep weather forecast models with high-resolution radar data. *PLOS Climate*, *4*(9), e0000633.

Kırboğa, K. K., Küçüksille, E. U., Naldan, M. E., Işık, M., Gülcü, O., & Aksakal, E. (2023). CVD22: Explainable artificial intelligence determination of the relationship of troponin to D-Dimer, mortality, and CK-MB in COVID-19 patients. *Computer Methods and Programs in Biomedicine*, *233*, 107492.

Klein, L., Ziegler, S., Laufer, F., Debus, C., Götz, M., Maier-Hein, K., ... Jäger, P. F. (2024). Discovering process dynamics for scalable perovskite solar cell manufacturing with explainable AI. *Advanced Materials*, *36*(7), 2307160.

Lötsch, J., & Ultsch, A. (2022). Enhancing explainable machine learning by reconsidering initially unselected items in feature selection for classification. *BioMedInformatics*, *2*(4), 701–714.

Lu, L., Zhu, T., Ribeiro, A. H., Clifton, L., Zhao, E., Zhou, J., ... Clifton, D. A. (2024). Decoding 2.3 million ECGs: interpretable deep learning for advancing cardiovascular diagnosis and mortality risk stratification. *European Heart Journal-Digital Health*, *5*(3), 247–259.

Maree, M. (2025). Quantifying relational exploration in cultural heritage knowledge graphs with LLMS: a neuro-symbolic approach for enhanced knowledge discovery. *Data*, *10*(4), 52.

Mühlberg, A., Ritter, P., Langer, S., Goossens, C., Nübler, S., Schneidereit, D., ... others (2023). SEMPAI: a self-enhancing multi-photon artificial intelligence for prior-informed assessment of muscle function and pathology. *Advanced Science*, *10*(28), 2206319.

Naser, M. (2022). Causality in structural engineering: Discovering new knowledge by tying induction and deduction via mapping functions and explainable artificial intelligence. *AI in Civil Engineering*, *1*(1), 6.

Papageorgiou, G., Sarlis, V., Maragoudakis, M., & Tjortjis, C. (2025). Hybrid multi-agent graphrag for e-government: Towards a trustworthy AI assistant. *Applied Sciences*, *15*(11), 6315.

PRISMA. (2024). *PRISMA statement.* Retrieved from `https://www.prisma-statement.org/`

Saluja, P., & Bhavsar, M. (2025). KDCFI AI-based knowledge discovery framework for cloud forensic investigation: P. saluja et al. *The Journal of Supercomputing*, *81*(17), 1572.

Shahrzad, H., Hodjat, B., & Miikkulainen, R. (2022). Evolving explainable rule sets. In *Proceedings of the genetic and evolutionary computation conference companion* (pp. 1779–1784).

Sharma, S., & Jain, S. (2024). OntoXAI: a semantic web rule language approach for explainable artificial intelligence. *Cluster Computing*, *27*(10), 14951–14975.

Shu, X., & Ye, Y. (2023). Knowledge Discovery: Methods from data mining and machine learning. *Social Science Research*, *110*, 102817.

Talukder, A. K., Sanz, J. B., & Samajpati, J. (2020). 'Precision health': Balancing reactive care and proactive care through the evidence based knowledge graph constructed from real-world electronic health records, disease trajectories, diseasome, and patholome. In *International conference on big data analytics* (pp. 113–133).

Tarnowska, K. A., Ras, Z. W., & Jastreboff, P. J. (2022). A data-driven approach to clinical decision support in tinnitus retraining therapy. *Frontiers in neuroinformatics*, *16*, 934433.

Thunold, H. H., Riegler, M. A., Yazidi, A., & Hammer, H. L. (2023). A deep diagnostic framework using explainable artificial intelligence and clustering. *Diagnostics*, *13*(22), 3413.

Vlahek, D., & Mongus, D. (2021). An efficient iterative approach to explainable feature learning. *IEEE transactions on neural networks and learning systems*, *34*(5), 2606–2618.

Waghen, K., & Ouali, M.-S. (2019). Interpretable logic tree analysis: A data-driven fault tree methodology for causality analysis. *Expert Systems with Applications*, *136*, 376–391.

Wagner, P., Mehari, T., Haverkamp, W., & Strodthoff, N. (2024). Explaining deep learning for ECG analysis: Building blocks for auditing and knowledge discovery. *Computers in biology and medicine*, *176*, 108525.

Wang, D., Chen, Y., & Chen, S. (2024). Discovering an interpretable mathematical expression for a full wind-turbine wake with artificial intelligence enhanced symbolic regression. *Physics of Fluids*, *36*(10).

Wang, X., Hosseininasab, A., Colunga, P., Kadıoğlu, S., & van Hoeve, W.-J. (2022). Seq2Pat: Sequence-to-pattern generation for constraint-based sequential pattern mining. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 36, pp. 12665–12671).

Withnell, E., Zhang, X., Sun, K., & Guo, Y. (2021). XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data. *Briefings in bioinformatics*, *22*(6), bbab315.

Xu, H., Chen, Y., Zeng, Z., Li, N., Li, J., & Zhang, D. (2025). Exploring terrain-precipitation relationships with interpretable AI for advancing future climate projections. *Nexus*, *2*(1).

Xu, L., Du, Z., Cai, M., Yin, S., Dong, S., Thanh, H. V., ... Dai, Z. (2025). Sparse data-driven knowledge discovery for interpretable prediction

of permeability in tight sandstones. *Engineering Geology*, *353*, 108151.

Xu, Y., Guo, H., Chen, S., & Liu, J. (2025). $\delta 1$: An automated theorem generator. *International Journal of Computational Intelligence Systems*, *18*(1), 309.

Zhe, Y., Hou, K., Wang, Z., Yang, S., & Sun, H. (2025). Prediction model for grouting volume using borehole image features and explainable artificial intelligence. *Construction and Building Materials*, *470*, 140626.

A Appendix

Index of Selected Documents

Title	Reference	Initialism / Abbreviation
Interpretable logic tree analysis: A data-driven fault tree methodology for causality analysis	Waghen & Ouali (2019)	Interpretable Logic Tree Analysis (ILTA)
'Precision health': Balancing reactive care and proactive care through the evidence based knowledge graph constructed from real-world electronic health records, disease trajectories, diseasome, and patholome	Talukder et al. (2020)	N/A
Introducing Uncertainty into Explainable AI Methods	Bobek & Nalepa (2021)	Local Uncertain Explanations (LUX)
XOmiVAE: an interpretable deep learning model for cancer classification using high-dimensional omics data	Withnell et al. (2021)	eXplainable Omics Variational AutoEncoder (XOmiVAE)
An Information Granulation Approach Through m-Grams for Text Classification	De Santis et al. (2021)	N/A
An Efficient Iterative Approach to Explainable Feature Learning	Vlahek & Mongus (2021)	N/A
Seq2Pat: Sequence-to-Pattern Generation for Constraint-Based Sequential Pattern Mining	X. Wang et al. (2022)	Sequence-to-Pattern (Seq2Pat)
Unsupervised explainable AI for molecular evolutionary study of forty thousand SARS-CoV-2 genomes	Iwasaki et al. (2022)	Batch-Learning Self-Organising Map (BLSOM)
Causality in structural engineering: discovering new knowledge by tying induction and deduction via mapping functions and explainable artificial intelligence	Naser (2022)	N/A
Supervised machine learning predictive analytics for alumni income	Gomez-Cravioto et al. (2022)	N/A
A data-driven approach to clinical decision support in tinnitus retraining therapy	Tarnowska et al. (2022)	Clinical Decision Support System (CDSS)
Evolving explainable rule sets	Shahrzad et al. (2022)	Evolution of Transparent Explainable Rule-sets (EVOTER)
Enhancing Explainable Machine Learning by Reconsidering Initially Unselected Items in Feature Selection for Classification	Lötsch & Ultsch (2022)	eXplainable Feature Selection (XFS)
Machine Learning Techniques for Prognosis Estimation and Knowledge Discovery From Lab Test Results With Application to the COVID-19 Emergency	Gerevini et al. (2023)	N/A
An Explainable Model to Support the Decision About the Therapy Protocol for AML	Almeida et al. (2023)	Explainable Boosting Machine (EBM)
SEMPAI: a Self-Enhancing Multi-Photon Artificial Intelligence for Prior-Informed Assessment of Muscle Function and Pathology	Mühlberg et al. (2023)	Self-Enhancing Multi-Photon Artificial Intelligence (SEMPAI)
CVD22: Explainable artificial intelligence determination of the relationship of troponin to D-Dimer, mortality, and CK-MB in COVID-19 patients	Kırboğa et al. (2023)	CardioVascular Diseases 22 (CVD22)

Title	Reference	Initialism / Abbreviation
A Deep Diagnostic Framework Using Explainable Artificial Intelligence and Clustering	Thunold et al. (2023)	N/A
XA-GANomaly: An Explainable Adaptive Semi-Supervised Learning Method for Intrusion Detection Using GANomaly	Han & Chang (2023)	eXplainable Adaptive-Generative Adversarial Network for anomaly detection (XA-GANomaly); t-distributed stochastic neighbour embedding (t-SNE)
Discovering an interpretable mathematical expression for a full wind-turbine wake with artificial intelligence enhanced symbolic regression	D. Wang et al. (2024)	Symbolic Regression (SR)
Neurosymbolic Learning in the XAI Framework for Enhanced Cyberattack Detection with Expert Knowledge Integration	Kalutharage et al. (2024)	Neuro-Symbolic (NS)
OntoXAI: a semantic web rule language approach for explainable artificial intelligence	Sharma & Jain (2024)	Ontonology eXplainable Artificial Intelligence (OntoXAI); Semantic Web Rule Language (SWRL)
Explaining deep learning for ECG analysis: Building blocks for auditing and knowledge discovery	Wagner et al. (2024)	N/A
CACTUS: A Comprehensive Abstraction and Classification Tool for Uncovering Structures	Gherardini et al. (2024)	Comprehensive Abstraction and Classification Tool for Uncovering Structures (CACTUS)
Discovering Process Dynamics for Scalable Perovskite Solar Cell Manufacturing with Explainable AI	Klein et al. (2024)	Counterfactual Explanations (CEs); Testing of Concept Activation Vectors (TCAV)
Decoding 2.3 million ECGs: interpretable deep learning for advancing cardiovascular diagnosis and mortality risk stratification	Lu et al. (2024)	N/A
AudioProtoPNet: An interpretable deep learning model for bird sound classification	Heinrich et al. (2025)	Audio Prototypical Part Network (AudioProtoPNet)
Quantifying Relational Exploration in Cultural Heritage Knowledge Graphs with LLMs: A Neuro-Symbolic Approach for Enhanced Knowledge Discovery	Maree (2025)	Neuro-Symbolic (NS)
KDCFI AI-based knowledge discovery framework for cloud forensic investigation	Saluja & Bhavsar (2025)	Knowledge Discovery Framework for Cloud Forensic Investigation (KDCFI)

Title	Reference	Initialism / Abbreviation
Prediction model for grouting volume using borehole image features and explainable artificial intelligence	Zhe et al. (2025)	N/A
Exploring terrain-precipitation relationships with interpretable AI for advancing future climate projections	H. Xu et al. (2025)	Genetic Algorithm-Geographic Weighted Regression (GA-GWR)
$\Delta 1$: An Automated Theorem Generator	Y. Xu et al. (2025)	Delta 1 ($\Delta 1$)
Unsupervised concept discovery for deep weather forecast models with high-resolution radar data	Kim et al. (2025)	N/A
Automatic labelling of topics in topic modelling	Angerri et al. (2025)	Principle Component Analysis (PCA)
Sparse data-driven knowledge discovery for interpretable prediction of permeability in tight sandstones	L. Xu et al. (2025)	N/A
Identifying lightning processes in ERA5 soundings with deep learning	Ehrensperger et al. (2025)	N/A
Interactive Discovery of Concept Drift with Lossless Visualization in Machine Learning	Gâlmeanu et al. (2025)	N/A
Hybrid Multi-Agent GraphRAG for E-Government: Towards a Trustworthy AI Assistant	Papageorgiou et al. (2025)	GraphRAG (Graph Retrieval Augmented Generation)
Concept drift detection and visualization with shifting window	Gâlmeanu & Andonie (2025)	N/A
Highway to.. Determining Fatal Outcomes in Traffic Accidents Based on Police Reports	Gabardo et al. (2025)	N/A
Patient complexity profiles in depression: a machine learning approach to personalized mental health	Dagnino et al. (2026)	Principle Component Analysis (PCA)
An explainable and adaptive Internet of Things intrusion detection system supported by Large Language Models	Huang et al. (2026)	Denoising Convolutional Autoencoder (DCAE)