

- ORIGINAL ARTICLE -

Explainable Artificial Intelligence: Analysis of Methodologies and Applications

Inteligencia Artificial Explicable: Análisis de Metodologías y Aplicaciones

María Cecilia Pezzini¹ , Claudia Pons^{1,2,3} ¹*Facultad de Informática, Universidad Nacional de La Plata, Argentina*²*Comisión de Investigaciones Científicas (CIC), Buenos Aires, Argentina*³*Facultad de Tecnología Informática, Universidad Abierta Interamericana, Argentina*

mcpazzini@gmail.com, cpons@lifa.info.unlp.edu.ar

Abstract

Explainability is essential in healthcare, finance, and security, where black-box models can undermine trust and decisions. Recent advances in eXplainable Artificial Intelligence (XAI) across structured/tabular data, computer vision, and natural language processing are surveyed. Thirty articles (2022–2024) were selected through a structured search with explicit inclusion criteria, and emerging approaches are compared with established techniques such as LIME and SHAP, alongside rule-, logic-, and ontology-based methods. Methods are organized along key dimensions—post-hoc vs. ante-hoc, model-agnostic vs. model-specific, scope, problem type, input data, and output format—and their effectiveness and applicability are evaluated. The review highlights innovations including spatially explainable architectures (e.g., SAMCNet) and entropy-based logic explanations, and identifies persistent challenges in robustness, cross-domain generalization, and deployment. Overall, findings consolidate the evolving XAI landscape and indicate directions toward reproducible techniques that strengthen transparency, accountability, and user trust in AI systems.

Keywords: artificial intelligence, explainability, explainable artificial intelligence, machine learning

Resumen

La explicabilidad es esencial en salud, finanzas y seguridad, donde los modelos de caja negra pueden socavar la confianza y las decisiones. Se revisan avances recientes en Inteligencia Artificial Explicable (XAI) en datos estructurados/tabulares, visión por computadora y procesamiento del lenguaje natural. Se seleccionaron treinta artículos (2022–2024) mediante una búsqueda estructurada con criterios de inclusión explícitos, y se comparan enfoques emergentes con técnicas consolidadas como LIME y SHAP, junto con métodos basados en reglas, lógica y ontologías. Los

métodos se organizan según dimensiones clave—post hoc vs. ante hoc, agnóstico al modelo vs. específico de modelo, alcance, tipo de problema, datos de entrada y formato de salida—y se evalúa su efectividad y aplicabilidad. Se destacan innovaciones como arquitecturas con explicabilidad espacial (p. ej., SAMCNet) y explicaciones lógicas basadas en entropía, y se identifican desafíos persistentes en robustez, generalización entre dominios y despliegue. En conjunto, los hallazgos consolidan el panorama en evolución de XAI y señalan direcciones hacia técnicas reproducibles que fortalezcan la transparencia, la rendición de cuentas y la confianza de los usuarios en los sistemas de IA.

Palabras claves: aprendizaje automático, explicabilidad, inteligencia artificial, inteligencia artificial explicable

1 Introduction

Artificial Intelligence (AI) has become a key factor across multiple domains, driving improvements in productivity, efficiency, and innovation [1, 2, 3]. Nevertheless, many systems still operate as opaque “black-box” models [4, 5, 6], which challenges trust, accountability, and acceptance, particularly in domains where automated decisions directly affect human lives, such as healthcare, finance, and security [7, 8, 9, 10, 11, 12].

In *The Fourth Industrial Revolution* [13], Schwab highlights the ethical and regulatory dilemmas stemming from the accelerated pace of technological change. He further stresses the need to establish robust regulatory frameworks and to foster collaboration among companies, governments, academia, and civil society to address these challenges effectively.

In this scenario, eXplainable Artificial Intelligence (XAI) plays a central role. Although no absolute consensus exists regarding the boundaries between *explainability* and *interpretability* [5, 6, 4], a pragmatic standpoint is adopted and detailed in Section 2.

Advances and challenges related to explainability span a wide range of application domains. In particular, this study examines models based on tabular data, computer vision, natural language processing, time series, medical imaging, and recommender systems. This article extends and refines a prior *final integrative project* available in the SEDICI repository [14]. While that work provided an initial synthesis of XAI methodologies, the present article expands the corpus to thirty recent contributions (2022–2024), introduces a refined taxonomy, and conducts a comparative analysis of trends and open challenges.

Accordingly, the following **research question** is addressed: “*What are the latest advances in improving the explainability of black-box machine learning models, and what is the impact of these advances in terms of proposals, taxonomies, and other relevant outcomes?*”

The main **contributions** of this work are as follows:

- A refined taxonomy of XAI methods organized by scenario, model specificity, scope, problem type, input data, and output format;
- A systematic selection and synthesis of thirty recent articles (2022–2024), with explicit counts by modality and technique;
- A comparative analysis of trends, open challenges, and opportunities for generalizable, model-agnostic explainability.

The remainder of the paper is organized as follows. Section 2 presents the conceptual framework; Section 3 reviews the literature; Section 4 analyzes XAI methodologies and applications—including Table 4-5, “Comparison of XAI methods: model analysis,” and Table 6, “Classification of XAI methods according to explainability criteria” [15]; and Section 6 presents conclusions and future work.

2 Conceptual Framework

This section defines neural networks and explainability in Artificial Intelligence (AI) models, providing a theoretical foundation. Special emphasis is placed on Artificial Neural Networks (ANNs), particularly Deep Neural Networks (DNNs). Categories of eXplainable Artificial Intelligence (XAI) and a taxonomy used to structure and analyze recent proposals are presented.

2.1 Concepts and Theoretical Background

2.1.1 Artificial and Deep Neural Networks

Artificial Neural Networks (ANNs) are inspired by the functioning of biological neurons. In ANNs, artificial neurons learn from data to perform tasks such as classification and prediction. A common type of ANN is the feedforward network, where information flows

from the input to the output layer without feedback or cycles [3].

Deep Neural Networks (DNNs), as an extension of feedforward ANNs, incorporate multiple hidden layers. This structure enables them to capture complex and nonlinear patterns, enhancing performance in applications such as computer vision and natural language processing [2].

2.1.2 Explainability in AI Models

XAI denotes methods and techniques that make the behavior and decisions of machine learning models understandable to humans. In this review, *interpretability* refers to intrinsic model properties that enable direct human understanding (e.g., sparse linear models, shallow trees), whereas *explainability* refers to external procedures that clarify a model’s decisions (e.g., post-hoc attributions or counterfactuals). This convention is used throughout the paper. Interpretability and explainability are not treated as equivalent terms in the text or tables, while acknowledging that some authors use them interchangeably.

Representative families of XAI methods include:

- **Local surrogate models:** techniques such as Local Interpretable Model-agnostic Explanations (LIME), which approximate local decision boundaries of a complex model with a simpler interpretable model [10].
- **Feature attribution methods:** SHapley Additive exPlanations (SHAP), which distributes predictions among input features using Shapley values [11].
- **Saliency/gradient-based approaches:** e.g., Gradient-weighted Class Activation Mapping (Grad-CAM) highlights relevant image regions by backpropagating gradients [12].
- **Rule- and logic-based approaches:** symbolic methods that extract human-readable rules or logical constraints [8].
- **Ontology/knowledge-graph-based approaches:** integration of ontologies and knowledge graphs to deliver semantically grounded explanations [16].

These methods serve as the baseline for comparison in Section 4, where recent proposals (2022–2024) are analyzed.

2.2 XAI Taxonomy

Explainability methods are classified according to scenario (ante-hoc/post-hoc), model specificity (agnostic/specific), scope (global/local), problem type (classification/regression), input data

Table 1: Key criteria and dimensions used to classify XAI methods.

Category	Subcategory
Scenario	Ante-hoc: explanations integrated during model development. Post-hoc: explanations generated after the model produces a prediction.
Model	Agnostic: applicable to any AI model. Specific: designed for a specific type of model.
Scope	Global: explanations covering the entire model. Local: explanations focused on specific instances.
Problem type	Classification: methods explaining classification models. Regression: methods explaining regression models.
Input data	Numeric/categorical: structured tabular data. Image: visual explanations such as heat maps. Text: highlighting important words or phrases. Time series: identifying patterns in sequential data.
Output format	Rules: logical rules. Numerical information: quantitative outputs. Textual information: written descriptions. Visual information: visualizations and images. Combination: mixed formats (e.g., text and visualizations).

(numeric/categorical, image, text, time series), and output format (rules, numerical, textual, visual, combination). Table 1 summarizes these criteria.

3 Literature Review

The systematic review follows the PICO model [17], summarized in Table 2. Keywords were used to guide the search for relevant literature addressing the research question: “*What are the latest advances in improving the explainability of black-box machine learning models, and what is the impact of these advances in terms of proposals, taxonomies, and other relevant outcomes?*”

3.1 Search Strategy

The search was conducted in major academic databases, including ACM Digital Library, IEEE Xplore, and Google Scholar. Queries combined the

Table 2: PICO approach by Kitchenham [17]

Aspect	Keywords / Query terms
Population	(“artificial intelligence” OR “machine learning” OR “deep learning”) AND (model OR system OR algorithm)
Intervention	(explanation OR explainability OR interpretable) AND (taxonomy OR method OR technique)
Comparison	(contrast OR differentiation)
Outcome	(advances OR proposals OR taxonomies)

Note. The term *interpretable* was included in the search query as a retrieval synonym to broaden coverage, since many papers use both terms interchangeably in titles or abstracts (see § 2.1.2 for the conceptual distinction).

PICO keywords and were applied to title, abstract, and keyword fields. To ensure recency, the review was limited to publications from January 2022 to May 2024, with priority given to open-access papers. This process yielded a broad set of candidate studies addressing explainability in black-box models.

3.2 Inclusion and Exclusion Criteria

To guarantee quality and relevance, studies were included that: (i) proposed advances in the explainability of black-box machine learning models; (ii) explicitly compared new techniques with previous approaches; and (iii) assessed the impact of these advances in terms of proposals, taxonomies, or other practical outcomes. Works were excluded if they lacked a clear focus on explainability, did not include methodological or comparative analysis, or were not openly accessible.

3.3 Article Selection and Analysis

Applying the above criteria resulted in a final corpus of 30 articles, as detailed in Section 4. Each paper was examined with respect to its methodological contributions, the problem type addressed (classification, regression, anomaly detection, etc.), the data modality considered (tabular, image, text, time series, medical imaging), the application domain (e.g., recommender systems), and the form of explanation produced (rules, attribution, visualization, textual description, hybrid).

3.4 Methodology for Analyzing Selected Articles

After applying the inclusion/exclusion criteria (Section 3.2), each paper was analyzed in five steps: (i) close reading to extract objectives and claimed contributions; (ii) classification by explainability

Table 3: Selected works on XAI.

ID	Title of the Work
1	Contrastive counterfactual fairness in algorithmic decision-making [18]
2	Federated explainability for network anomaly characterization [19]
3	RoCourseNet: robust training of a prediction-aware recourse model [20]
4	SAMCNet: Towards a Spatially Explainable AI Approach for Classifying MxIF Oncology Data [21]
5	Subgoal-based explanations for unreliable intelligent decision support systems [22]
6	LCNN: Lightweight CNN architecture for software defect feature identification using Explainable AI [23]
7	Deep prototypical-parts ease morphological kidney stone identification and are competitively robust to photometric perturbations [24]
8	Fuzzy rule-based explainer systems for deep neural networks: from local explainability to global understanding [25]
9	Explainable artificial intelligence: what we know and what is left to attain trustworthy artificial intelligence [9]
10	Entropy-based logic explanations of neural networks [26]
11	Logic Explained Networks (LENs) [27]
12	Deep learning with logical constraints [28]
13	Robust explainability: a tutorial on gradient-based attribution methods for deep neural networks [29]
14	Explaining the black-box smoothly — a counterfactual approach [30]
15	Post-hoc Concept Bottleneck Models [31]
16	sMRI-PatchNet: explainable patch-based deep learning for Alzheimer’s diagnosis with structural MRI [32]
17	Supervised contrastive learning for interpretable long-form document matching [33]
18	A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models (Q-MCKT) [34]
19	Identifying explanation needs of end-users: applying and extending the XAI Question Bank [35]
20	Approaching XAI methods in the diagnosis of iron deficiency anemia using blood parameters [36]
21	XAI for medicine by ChatGPT code interpreter [37]
22	Interpreting black-box machine learning models for high-dimensional datasets [38]
23	xAI: an explainable AI model for the diagnosis of COPD from CXR images [39]
24	Explainable artificial intelligence and cardiac imaging: toward more interpretable models [40]
25	A spectrum of explainable and interpretable machine learning approaches for genomic studies [41]
26	A survey on medical explainable AI (XAI): recent progress, approaches, human interaction and scoring system [42]
27	Knowledge graphs as tools for explainable machine learning: a survey [16]
28	Interpretable local concept-based explanation with human feedback to predict all-cause mortality [43]
29	On the explainability of natural language processing deep models [44]
30	A comprehensive review and application of interpretable deep learning models for ADR prediction [45]

technique and evaluation setting; (iii) comparative mapping against prior approaches (baseline/ablation, datasets, metrics); (iv) assessment of reported impact (proposals, taxonomies/frameworks, practical use cases); (v) synthesis into cross-study tables.

4 Analysis of XAI Methodologies and Applications

This section describes the corpus that met the inclusion criteria and structures its attributes for subsequent analysis. Table 3 summarizes the thirty works identified by the systematic search in Section 3. Each item was evaluated for relevance and contribution to XAI regarding interpretability, robustness to perturbations, and cross-context applicability. The resulting dataset underpins the comparative study presented in Tables 4–5 and the taxonomy-based classification in Table 6.

4.1 Explanation of the columns in Tables 4 and 5

The analysis dimensions derive from the review’s research question; the columns are defined as follows:

- **Method / Approach Used** — Explainability family or concrete technique (e.g., LIME,

SHAP, Grad-CAM, rule- or logic-based, ontology/knowledge-graph-based).

- **Proposal** — Main contribution or objective of the method (e.g., new attribution mechanism, counterfactual generation, symbolic constraint).
- **Area of Use** — Application domain (e.g., healthcare, fraud detection, image recognition, recommender systems).
- **Main Benefits** — Reported advantages (e.g., improved interpretability, reduced complexity, human-centered outputs).
- **Addressed Problem** — Targeted task or challenge (e.g., understanding black-box models, interpreting DNNs, real-time explanations).
- **Generalizability** — Indicates whether the explainability approach can be broadly applied across different black-box models.
- **Data Transformation** — Required preprocessing or transformations (e.g., normalization, feature extraction, segmentation).
- **Examples** — Representative use cases or prior studies where the method has been applied.

Table 4: Comparison of XAI methods: model analysis (Part A). Columns ID to Addressed Problem.

ID	Method / Approach Used	Proposal / Objective	Area of Use (Application Domain)	Main Benefits	Addressed Problem
1	CF Contrastive fairness	Contrastive to improve fairness.	Loans; hiring; credit.	transparency/trust.	Fairness CF explanations.
2	FL for intrusion detection	Explain anomalies in unsupervised FL IDS.	Distributed IoT NIDS.	interpretability/efficacy.	Anomaly explanations.
3	RoCourseNet (tri-level, adv., VDS)	Robust CFs via VDS/CounterNet.	CF under distribution shift.	Valid/robust; transparency; generalization.	CFs under shift.
4	Spatially Aware Multi-category Convolutional Neural Network (SAMCNet)	Asymmetric aggregation of spatial features.	Oncology; pharma; biomed; paleo; ecology; epidemiology.	accuracy; efficiency; pattern discovery.	Explain spatial patterns.
5	Subgoal-based explanations	Subgoal explanations for IDS.	Decision support (IDS).	user performance; robust to failures.	Explain recommendations.
6	LIME + SHAP	Interpret CNN for defect prediction.	Software defects.	efficiency/accuracy; local+global views.	Model explanations.
7	Prototypical parts / ProtoPNet	Hierarchical prototypes.	CADx: kidney stones.	accuracy/explainability; clinician confidence.	Model explanations.
8	Fuzzy rule-based explainers	Fuzzy rules to explain DNNs.	Medicine; finance; applied sci.	Interpretability; fidelity; fewer rules/features.	Model explanations.
9	XAI Methodology (Four Axes)	Framework: data, model, post-hoc, eval.	DS; devs; experts; end users.	Transparency; alignment; design support; trust.	XAI design guidance.
10	Entropy-based logic explanations	Entropy criterion for explanations.	First-order logic tasks.	Formal; concise.	Logic explanations.
11	LENs	First-order logic for explanations.	Health; finance; science; education.	Flexible granularity; interpretability.	Rule extraction; local/global expl.
12	Logical constraints	Embed constraints in loss/outputs.	Broad domains.	Domain rules; interpretability; robustness.	Constraint-aware design.
13	Gradient-based attribution	Survey + robustness evaluation.	Medical AI; auto-driving; critical.	Fast; robustness/decision-load eval.	Attribution explanations.
14	GAN CF explainer	CFs preserving details/metrics.	Medical imaging.	Transparency; clinical detail; benchmarking.	Counterfactual explanations.
15	P-CBM / HP-CBM	Post-hoc/hybrid CBMs with residuals.	Medical diagnosis.	Versatile; no concept labels; gains w/o tuning.	Concept/local explanations.
16	sMRI-PatchNet	Auto patch selection; ROI; comparisons.	Medical diagnosis.	accuracy; faster; generalization.	Model explanations.
17	Contrastive Long-Document Encoding (CoLDE)	Contrastive learning + chunkwise MHA; hierarchical docs.	Research; patents; retrieval; long docs.	Detailed; strong XAI; high perf.; scalable.	Model explanations.
18	Q-MCKT	Question-centric KT; experts; contrastive; IRT.	Education (KT).	accuracy; learning; interpretable; robust imbalance.	Model explanations.
19	XAIQB (extended)	More questions/descriptions for users.	End users.	Personalized; transparency; better interaction/decisions.	User-centered needs.
20	XAI for iron-deficiency anemia	Classifier + beeswarm for blood counts.	Medical diagnosis.	Efficient; transparent; trusted; attribute-level.	Model explanations.
21	ChatGPT Code Interpreter + PBC	Prompt-based code for medical texts.	Medical diagnosis.	explainability; MAPC compliance.	Model explanations.
22	Surrogates + FI + perturbations	Interpretable surrogates via feature ID/perturbation.	CV; NLP; tabular.	XAI in high-dimensional data.	Model explanations.
23	Grad-CAM + SHAP (COPD)	Early COPD from X-rays.	COPD diagnosis.	Early/accurate; trust; low-resource viable.	Model explanations.
24	Grad-CAM + SHAP + LIME + SmoothGrad	Open-source tools; app/human/function evals.	Cardiac imaging.	Understandable results; transparency/clarity.	Model explanations.
25	SHAP + LIME + bio knowledge	PCNNs + bio-losses for genomics ML.	Genomics.	Interpretability; stronger bio grounding.	Model expl./interpretability.
26	LIME; Grad-CAM; XAI-RS; XAI-SS	User-centered XAI: collaboration/feedback/eval/scoring.	Medical diagnosis.	Transparency; legal/ethical; R&D; collaboration.	Model explanations.
27	KBX-systems	Symbolic/KB XAI; data integration + interaction.	Recommenders; image/item; mining; conversational.	Symbolic expl.; reuse/adaptation; context; learning.	Model explanations.
28	Concept-based Local Explanations with Feedback (CLEF)	Concept-based local expl. + expert feedback + CFs.	Clinical diagnosis.	Clear; collaboration; CF insights; bias detect; fidelity.	Local/result/CF explanations.
29	exBERT; ERASER (NLP)	NLP XAI tools/benchmarks.	NLP.	interpretability; method taxonomy.	Model explanations.
30	SP-LIME (ADRs)	Interpretable ADR prediction.	Medical ADR.	risk; optimized treatment.	Model explanations.

• Comparison with Previous Approaches — Key differences versus earlier explainability techniques.

Index by method (IDs).

- **Counterfactual methods:** 1, 3, 14
- **Gradient-/attribution-/surrogate-based:** 6, 13, 22, 23, 24, 30
- **Logic-/rule-/ontology-/knowledge-based:** 8, 10, 11, 12, 25, 27
- **Prototype-/concept-based:** 7, 15, 16, 28
- **Methodology/frameworks/user-centered:** 5, 9, 19, 26
- **Federated/tools/platforms:** 2, 21
- **Other application-focused (domain-specific):** 4, 17, 18, 20, 29

Reader guide. IDs match Table 3 (titles) and Tables 4–6 (column-wise analysis). Rows remain in ID order (1–30) to preserve one-to-one traceability across tables and figures.

4.2 Discussion and Analysis of Trends in Explainability

The following trends are observed (IDs refer to Tables 4 and 5):

- 1) **Counterfactual-based approaches.** Generation of counterfactual explanations that show how decisions could change under alternative conditions (IDs 1, 3, 14).
- 2) **Interactive models and personalized explanations.** Approaches that adapt explanations to user needs, such as interpretable long-document matching and post-hoc concept bottlenecks (IDs 17, 15).
- 3) **Integration of external knowledge and logic.**

Table 5: Comparison of XAI methods: model analysis (Part B). Columns ID and Generalizability to Comparison with Previous Approaches.

ID	Generalizability	Data Transformation	Examples	Comparison with Previous Approaches
1	YES	NO	NO	Provides fairness explanations without requiring mitigation strategies or extra CF data.
2	YES	YES	YES	Uses SHAP and LEMNA as explanatory tools.
3	YES	YES	YES	Compared with CounterNet, ROAR and RB in robustness settings.
4	YES	YES	YES	Benchmarked against PointNet, DGCNN and SRNet.
5	YES	NO	YES	Compared with aIDS, Action Recommendation from IDS, and EC-LC (Causal-Link Chain Explanation).
6	YES	NO	YES	Compared with Deep Representation & Ensemble Learning, DAECNN-JDP, and transfer CNN models.
7	YES	NO	YES	Compared with base CNNs in accuracy, mean precision, and F1 score under IID/OOD conditions.
8	YES	YES	YES	Compared with: FRBCS; and ECLAIRE.
9	YES	NO	YES	
10	YES	NO	YES	Compared with ψ Networks; Decision Trees; and Bayesian Rule Lists.
11	YES	YES	YES	Compared with LIME; SHAP; AM; Saliency Maps; SP-LIME; Decision Trees; BRL; and Deep-RED.
12	YES	NO	YES	Compared with KBANN; CILP; LTN; NeurASP; and Iterative Rule Distillation.
13	YES	NO	YES	Compared with Guided Backpropagation; Saliency Maps; DeepLIFT; Backpropagation with Grad-CAM; DeconvNet; and SmoothGrad.
14	YES	YES	YES	Compared with Saliency Maps and CycleGAN.
15	YES	NO	YES	Compared with traditional CBMs.
16	YES	NO	YES	Compared with SVM, LDA, and KNN.
17	YES	YES	YES	Compared with DSSM, ARC-I, HAN, Siamese-BERT, SBERT, SMITH, and S-LONG.
18	NO	YES	YES	Compared with LIME; SHAP; Grad-CAM; Integrated Gradients; Anchors; and counterfactual explanations.
19	YES	NO	YES	
20	YES	NO	YES	Compared with Logistic Regression; Random Forest; SVM; and KNN.
21	YES	YES	YES	Compared with text-based prompting (TBP).
22	YES	YES	YES	Compared with TabNet, XGBoost, and SHAP.
23	NO	YES	YES	Compared with ResNet50, Xception, Grad-CAM, and SHAP.
24	YES	YES	YES	Compared with classical ML models and deep learning models.
25	YES	YES	YES	Compared with traditional black-box models and conventional post-hoc techniques.
26	YES	YES	YES	Compared with SHAP and LRP.
27	YES	YES	YES	Compared with post-hoc methods: LIME; SHAP; rule-based; and decomposition approaches.
28	NO	NO	YES	Compared with interactive baseline (AL) and non-interactive baseline (LR).
29	YES	YES	YES	Compared with LIME and SHAP.
30	YES	YES	YES	Compared with SVM, Decision Trees, and KNN.

Table 6: Classification of XAI methods [15].

ID	XAI Scenario	XAI Scope	XAI Problem Type	XAI Input Data	XAI Output Format
1	Post-hoc (model-agnostic)	Global	Classification	Numeric / Categorical	Textual + Numerical
2	Post-hoc (model-agnostic)	Local, Global	Classification	Numeric / Categorical	Rules; Visual + Numerical
3	Post-hoc (model-agnostic)	Global	Classification	Numeric / Categorical; Text	Textual + Numerical
4	Post-hoc (model-specific)	Global	Classification	Image	Visual
5	Post-hoc (model-specific)	Local	Classification	Numeric / Categorical	Textual
6	Post-hoc (model-specific)	Local	Classification	Numeric / Categorical	Textual + Numerical
7	Post-hoc (model-specific)	Local	Classification	Image	Visual + Numerical
8	Post-hoc (model-specific)	Local, Global	Classification	Numeric / Categorical	Fuzzy rules
9	Post-hoc (model-agnostic)	Local	Classification, Regression	Numeric / Categorical; Text; Visual	Feature explanations; Visualizations
10	Post-hoc (model-agnostic)	Global	Classification	Numeric / Categorical; Image; Text	Visual + Numerical
11	Post-hoc (model-agnostic)	Global, Local	Classification	Numeric / Categorical	Rules + Numerical
12	Ante-hoc (model-agnostic)	Global, Local	Classification	Numeric / Categorical; Image	Visual + Numerical
13	Post-hoc (model-agnostic)	Global, Local	Classification	Image	Visual
14	Post-hoc (model-agnostic)	Global	Classification	Image	Visual
15	Post-hoc (model-agnostic)	Global	Classification	Numeric / Categorical; Image; Text; Time series	Rules; Textual; Visual; Numerical
16	Post-hoc (model-specific)	Global	Classification	Image	Visual
17	Ante-hoc	Global, Local	Classification	Text	Textual
18	Post-hoc (model-agnostic)	Local	Classification	Numeric / Categorical	Textual + Numerical
19	Ante-hoc	Global, Local	Classification, Regression	Numeric / Categorical; Image; Text; Time series	Mixed formats
20	Post-hoc (model-agnostic)	Local	Classification	Numeric / Categorical	Visual + Numerical
21	Ante-hoc / Post-hoc (model-specific)	Global, Local	Classification, Regression	Numeric / Categorical; Image; Text; Time series	Rules; Visual; Textual; Numerical
22	Post-hoc (model-agnostic)	Global, Local	Classification, Regression	Numeric / Categorical	Numerical
23	Post-hoc (model-agnostic)	Global, Local	Classification	Image	Visual
24	Post-hoc (model-agnostic / model-specific)	Global, Local	Classification, Regression	Numeric / Categorical; Image	Visual; Textual; Numerical
25	Ante-hoc / Post-hoc (model-specific)	Global, Local	Classification	Numeric / Categorical	Numerical
26	Post-hoc (model-agnostic)	Global	Classification	Numeric / Categorical; Image; Text; Time series	Rules; Visual; Textual; Numerical
27	Post-hoc (model-agnostic)	Global	Classification	Numeric / Categorical; Text	Rules; Visual; Textual; Numerical
28	Post-hoc (agnostic)	Local	Classif.	Num./Cat.	Textual; CF; Visual
29	Post-hoc (model-agnostic)	Global (ERASER), Local (exBERT)	Classification	Text	Visual; Textual
30	Post-hoc (model-agnostic)	Local	Classification	Numeric / Categorical	Textual

Logic Explained Networks and logical constraints to align explanations with domain knowledge and improve auditability (IDs 11, 12; see also 10, 25, 27).

- 4) **Explainability in distributed and federated models.** Techniques tailored to decentralized settings (e.g., federated learning) addressing interpretability and cross-site comparability (IDs 2, 21).
- 5) **Advances in visualization- and attribution-based methods.** Use of Grad-CAM, SHAP, and related pipelines to clarify feature influence, especially in imaging and clinical contexts (IDs 23, 24; robustness/tutorial guidance in ID 13; broader variants in 22, 30).
- 6) **Tools and benchmarking.** Tooling and benchmarks that support evaluation and comparison across tasks and domains (ID 29).

5 Classification of XAI Methods According to Explainability Criteria

Following Vilone and Longo's framework, XAI methods are classified along five dimensions—scenario, scope, problem type, input data, and output format—as summarized in Table 1 [15]. Using Table 6 as the source, Figs. 1–3 quantify the distribution of the 30 studies across (a) *problem type*, (b) *scope*, and (c) *explanation scenario*, respectively. These figures complement the tabular classification by highlighting aggregate trends without reiterating methodological definitions.

5.1 Explanation Scenarios

1. **Post-hoc (76.7%):** Most methods provide explanations after the model has been trained (e.g., Contrastive Counterfactual Fairness, Deep Prototypical-Parts). These methods aim to interpret decisions already made by the model.
2. **Ante-hoc (10.0%):** Less common, these methods integrate explanations into model training (e.g., Deep Learning with Logical Constraints),

improving transparency from the beginning of the training process.

3. **Hybrid (13.3%):** Some methods combine post-hoc and ante-hoc approaches, offering flexibility in the application of explainability techniques.

Scope of Explanations.

1. **Global (43.3%):** These methods provide global explanations, offering a comprehensive view of the model (e.g., Logic Explained Networks).
2. **Local (30.0%):** These methods focus on local explanations, detailing specific decisions (e.g., LCNN).
3. **Both (26.7%):** Some methods address both global and local explanations, providing an integrated view of the model (e.g., Fuzzy Rule-Based Explainer Systems).

Problem Type.

1. **Classification (63.3%):** The majority of the methods specialize in classification problems (e.g., SAMCNet, sMRI-PatchNet).
2. **Classification and Regression (36.7%):** A substantial share of methods address both classification and regression, although regression tasks receive comparatively less attention.

Of the 30 articles reviewed, 23 adopt post-hoc strategies, 3 develop ante-hoc approaches, and 4 combine both perspectives. Regarding the scope of explanations, 13 papers provide global explanations, 9 focus on local explanations, and 8 integrate both levels. Concerning problem type, 19 contributions address classification tasks exclusively, while 11 cover both classification and regression. These distributions confirm that post-hoc techniques remain the dominant paradigm, although interest in hybrid and ante-hoc strategies is gradually increasing. They also highlight the strong prevalence of classification and image-based tasks, with comparatively fewer works exploring regression or multimodal contexts.

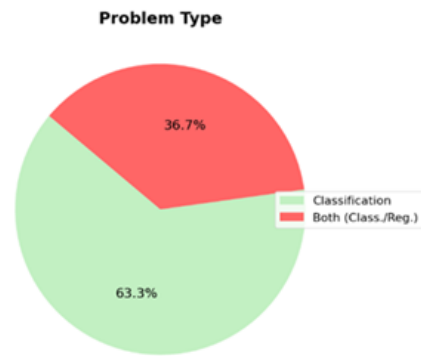


Figure 1: Problem type: proportion of methods addressing classification vs. regression.



Figure 2: Scope of explanations: share of global, local, and combined approaches.

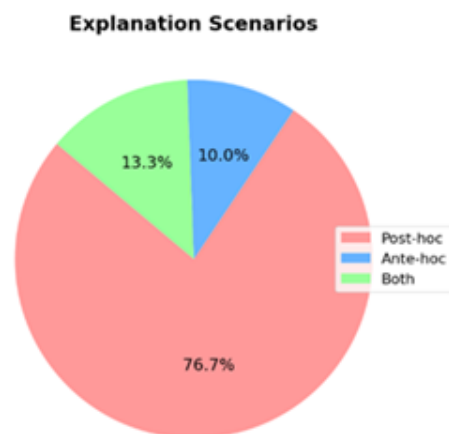


Figure 3: Explanation scenarios: prevalence of post-hoc, ante-hoc, and hybrid strategies.

6 Conclusions

This review addressed the research question: “*What are the latest advances in improving the explainability of black-box machine learning models, and what is the impact of these advances in terms of proposals, taxonomies, and other relevant outcomes?*” A systematic analysis of thirty recent contributions (2022–2024) identifies both consolidation of established techniques and novel directions that broaden the scope of XAI.

6.1 Advances in Explainability Techniques

The reviewed literature presents a wide spectrum of methods, from rule- and logic-based approaches to recent proposals such as Deep Prototypical Parts (PPs) and entropy-based criteria. This diversity reflects the dynamism of the field and the need for multiple strategies to enhance interpretability.

Among the most relevant advances, SAMCNet improves the visualization of spatial patterns in oncological data, while PPs facilitate the identification of salient features in complex models. In parallel, model-agnostic methods such as LIME and SHAP remain widely adopted due to their flexibility and applicability across diverse models, despite limitations related to perturbation-based explanations.

Other contributions, such as Entropy-Based Logic Explanations and Gradient-Based Attribution, integrate formal and mathematical principles to provide more rigorous explanations. These approaches highlight an emerging trend toward developing robust and generalizable techniques. Furthermore, several studies emphasize the inclusion of illustrative examples and comparisons with previous approaches, which strengthen the credibility and practical relevance of the proposed methods.

6.2 Challenges

Despite these advances, important challenges persist. Developing model-agnostic techniques that can be reliably applied across heterogeneous architectures remains difficult. Issues such as scalability, robustness to adversarial settings, and preservation of data integrity are still unresolved. Additionally, the lack of standardized benchmarks and evaluation metrics complicates systematic comparison between methods. Addressing these challenges is essential for the widespread adoption of XAI in real-world contexts.

6.3 Future Work

Future work in the field of eXplainable Artificial Intelligence should focus on evaluating the practical impact of explainability techniques in critical sectors, considering both technical performance and user acceptance. It will be essential

to tailor explanations to the specific needs of different stakeholders—clinicians, financial analysts, or regulators—to ensure transparency, fairness, and respect for privacy. Furthermore, the exploration of hybrid approaches that combine symbolic reasoning, visualization, and counterfactuals may open new directions toward more universal and trustworthy explainability frameworks.

This study constitutes a preliminary contribution that refines taxonomies, synthesizes recent methods, and identifies trends and gaps that may guide future investigations by the community and subsequent developments in the area.

Authors’ contribution

MC: Conceptualization, Methodology, Investigation, Analysis, and Writing – original draft. CP: Supervision, Writing – review & editing. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare that they have no competing interests.

Acronyms

Acronyms

ADR Adverse Drug Reaction.

AI Artificial Intelligence.

AL Active Learning.

AM Activation Maximization.

ANN Artificial Neural Network.

BRL Bayesian Rule Lists.

CADx Computer-Aided Diagnosis.

CLEF Concept-based Local Explanations with Feedback.

CNN Convolutional Neural Network.

CoLDE Contrastive Long-Document Encoding.

DAECNN Dual-Attention Enhanced CNN.

Deep-RED Deep Reverse Engineering of Deep Networks.

DNN Deep Neural Network.

DSSM Deep Structured Semantic Model.

ECLAIRE Explainable Classifier with Rule Extraction.

ERASER Evaluating Rationales And Simple English Reasoning.

FRBCS Fuzzy Rule-Based Classification System.

Grad-CAM Gradient-weighted Class Activation Mapping.

HAN Hierarchical Attention Network.

IDS Intelligent Decision Support Systems.

IID Independent and Identically Distributed.

KNN k-Nearest Neighbors.

LDA Linear Discriminant Analysis.

LENS Logic Explained Networks.
LIME Local Interpretable Model-agnostic Explanations.
LR Logistic Regression.
LRP Layer-wise Relevance Propagation.
LTN Logic Tensor Network.

ML Machine Learning.

NeurASP Neural Answer Set Programming.

OOD Out-of-Distribution.

ProtoPNet Prototypical Part Network.

S-LONG Structured LONG-document model.

SAMCNet Spatially Aware Multi-category Convolutional Neural Network.

SBERT Sentence-BERT.

SHAP SHapley Additive exPlanations.

SMITH Segmented Masked Language Model for Document-Level Tasks.

SVM Support Vector Machine.

XAI eXplainable Artificial Intelligence.

References

- [1] R. Baeza-Yates, “Introduction to responsible AI,” in *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM '24)*, 2024, pp. 1114–1117. [Online]. Available: <https://doi.org/10.1145/3616855.3636455>
- [2] I. Goodfellow, Y. Bengio, and A. Courville, *Deep learning*. Cambridge, MA, USA: MIT Press, 2017.
- [3] Y. LeCun, “Generalization and network design strategies,” University of Toronto, Department of Computer Science, Tech. Rep. CRG-TR-89-4, 1989.
- [4] C. Molnar, *Interpretable machine learning*, 2nd ed. Lulu.com, 2020.
- [5] F. Doshi-Velez and B. Kim, “Towards a rigorous science of interpretable machine learning,” 2017.
- [6] Z. C. Lipton, “The mythos of model interpretability,” 2016.
- [7] O. Biran and C. Cotton, “Explanation and justification in machine learning: A survey,” in *Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI Workshop on Explainable Artificial Intelligence)*, 2017.
- [8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, “A survey of methods for explaining black box models,” *ACM Computing Surveys*, vol. 51, no. 5, pp. 93:1–93:42, 2018. [Online]. Available: <https://doi.org/10.1145/3236009>
- [9] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, “Explainable artificial intelligence (XAI): What we know and what is left to attain trustworthy artificial intelligence,” *Information Fusion*, vol. 99, p. 101805, 2023. [Online]. Available: <https://doi.org/10.1016/j.inffus.2023.101805>
- [10] M. T. Ribeiro, S. Singh, and C. Guestrin, ““why should i trust you?” explaining the predictions of any classifier,” in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [11] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of Advances in Neural Information Processing Systems 30 (NeurIPS 2017)*, 2017, pp. 4765–4774.
- [12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, “Grad-CAM: Visual explanations from deep networks via gradient-based localization,” in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 618–626. [Online]. Available: <https://doi.org/10.1109/ICCV.2017.74>
- [13] K. Schwab, *The fourth industrial revolution*. New York, NY, USA: Crown Business, 2017.
- [14] M. C. Pezzini, “Inteligencia artificial explicable: Análisis de metodologías y aplicaciones,” 2024. [Online]. Available: <http://sedici.unlp.edu.ar/handle/10915/174328>
- [15] G. Vilone and L. Longo, “Notions of explainability and evaluation approaches for explainable artificial intelligence,” *Information Fusion*, vol. 76, pp. 89–106, 2021. [Online]. Available: <https://doi.org/10.1016/j.inffus.2021.05.009>
- [16] I. Tiddi and S. Schlobach, “Knowledge graphs as tools for explainable machine learning: A survey,” *Artificial Intelligence*, vol. 302, p. 103627, 2022. [Online]. Available: <https://doi.org/10.1016/j.artint.2021.103627>

- [17] B. A. Kitchenham and S. Charters, “Guidelines for performing systematic literature reviews in software engineering,” Keele University and Durham University, Tech. Rep. EBSE-2007-01, 2007.
- [18] E. Ç. Mutlu, N. Yousefi, and O. O. Garibay, “Contrastive counterfactual fairness in algorithmic decision-making,” in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES '22)*, 2022, pp. 499–507. [Online]. Available: <https://doi.org/10.1145/3514094.3534143>
- [19] X. Sáez-de Cámara, J. L. Flores, C. Arellano, A. Urbieto, and U. Zurutuza, “Federated explainability for network anomaly characterization,” in *Proceedings of the 26th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '23)*, 2023, pp. 346–365. [Online]. Available: <https://doi.org/10.1145/3607199.3607234>
- [20] H. Guo, F. Jia, J. Chen, A. Squicciarini, and A. Yadav, “Rocoursenet: Robust training of a prediction aware recourse model,” in *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management (CIKM '23)*, 2023, pp. 619–628. [Online]. Available: <https://doi.org/10.1145/3583780.3615040>
- [21] M. Farhadloo, C. Molnar, G. Luo, Y. Li, S. Shekhar, R. L. Maus, S. Markovic, A. Leontovich, and R. Moore, “Samcnet: Towards a spatially explainable ai approach for classifying mxif oncology data,” in *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22)*, 2022, pp. 2860–2870. [Online]. Available: <https://doi.org/10.1145/3534678.3539168>
- [22] D. Das, B. Kim, and S. Chernova, “Subgoal-based explanations for unreliable intelligent decision support systems,” in *Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23)*, 2023, pp. 240–250. [Online]. Available: <https://doi.org/10.1145/3581641.3584055>
- [23] M. Begum, M. H. Shuvo, M. K. Nasir, A. Hossain, M. J. Hossain, I. Ashraf, J. Uddin, and M. Samad, “Lcnn: Lightweight CNN architecture for software defect feature identification using explainable AI,” *IEEE Access*, vol. 12, pp. 55 744–55 756, 2024. [Online]. Available: <https://doi.org/10.1109/ACCESS.2024.3388489>
- [24] D. Flores-Araiza, F. Lopez-Tiro, E. Villalvazo-Avila, J. El-Beze, J. Hubert, G. Ochoa-Ruiz, and C. Daul, “Deep prototypical-parts ease morphological kidney stone identification and are competitively robust to photometric perturbations,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2023, pp. 295–304. [Online]. Available: <https://doi.org/10.1109/CVPRW59228.2023.00035>
- [25] F. Aghaeipoor, M. Sabokrou, and A. Fernández, “Fuzzy rule-based explainer systems for deep neural networks: From local explainability to global understanding,” *IEEE Transactions on Fuzzy Systems*, pp. 1–12, 2023. [Online]. Available: <https://doi.org/10.1109/TFUZZ.2023.3243935>
- [26] P. Barbiero, J. D. T. M. Silva, J. Zarlenga, Y. Qi *et al.*, “Entropy-based logic explanations of neural networks,” in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022, pp. 5935–5943. [Online]. Available: <https://doi.org/10.1609/aaai.v36i5.20345>
- [27] G. Ciravegna, P. Barbiero, F. Giannini, M. Gori, P. Liò, M. Maggini, and S. Melacci, “Logic explained networks,” *Artificial Intelligence*, vol. 314, p. 103822, 2023. [Online]. Available: <https://doi.org/10.1016/j.artint.2022.103822>
- [28] E. Giunchiglia, M. C. Stoian, and T. Lukasiewicz, “Deep learning with logical constraints,” 2022.
- [29] I. E. Nielsen, D. Dera, G. Rasool, R. P. Ramachandran, and N. C. Bouaynaya, “Robust explainability: A tutorial on gradient-based attribution methods for deep neural networks,” *IEEE Signal Processing Magazine*, vol. 39, no. 4, pp. 73–84, 2022. [Online]. Available: <https://doi.org/10.1109/MSP.2022.3142719>
- [30] S. Singla, M. Eslami, B. Pollack, S. Wallace, and K. Batmanghelich, “Explaining the black-box smoothly—a counterfactual approach,” *Medical Image Analysis*, vol. 84, p. 102721, 2023. [Online]. Available: <https://doi.org/10.1016/j.media.2022.102721>
- [31] M. Yuksekgonul, M. Wang, and J. Zou, “Post-hoc concept bottleneck models,” in *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023, arXiv:2205.15480.
- [32] X. Zhang, L. Han, L. Han, H. Chen, D. Dancey, and D. Zhang, “smri-patchnet: A novel efficient explainable patch-based deep learning network for alzheimer’s disease diagnosis with structural MRI,” *IEEE Access*, vol. 11, pp. 108 603–108 616, 2023. [Online]. Available: <https://doi.org/10.1109/ACCESS.2023.3321220>

- [33] A. Jha, V. Rakesh, J. Chandrashekar, A. Samavedhi, and C. K. Reddy, "Supervised contrastive learning for interpretable long-form document matching," *ACM Transactions on Knowledge Discovery from Data*, vol. 17, no. 2, p. Article 27, 2023. [Online]. Available: <https://doi.org/10.1145/3542822>
- [34] H. Zhang, N. Zeng, Y. Liu, Y. Zhang, and S. Chen, "A question-centric multi-experts contrastive learning framework for improving the accuracy and interpretability of deep sequential knowledge tracing models," *ACM Transactions on Intelligent Systems and Technology*, 2024. [Online]. Available: <https://doi.org/10.1145/3674840>
- [35] L. Sipos, U. Schäfer, K. Glinka, and C. Müller-Birn, "Identifying explanation needs of end-users: Applying and extending the XAI question bank," in *Proceedings of Mensch und Computer 2023 (MuC '23)*. Rapperswil, Switzerland: ACM, 2023, pp. 1–6. [Online]. Available: <https://doi.org/10.1145/3603555.360851>
- [36] U. Ponnusamy, D. D. B. S., and N. Sampathila, "Approaching explainable artificial intelligence methods in the diagnosis of iron deficiency anemia using blood parameters," in *Proceedings of the 2023 International Conference on Recent Advances in Information Technology for Sustainable Development (ICRAIS)*, 2023, pp. 201–206. [Online]. Available: <https://doi.org/10.1109/ICRAIS59684.2023.10367126>
- [37] K. Kitamura, M. Irvan, and R. S. Yamaguchi, "XAI for medicine by chatgpt code interpreter," in *Proceedings of the 5th International Conference on Big Data Service and Intelligent Computation (BDSIC 2023)*, Singapore, 2023, pp. 28–34. [Online]. Available: <https://doi.org/10.1145/3633624.3633629>
- [38] M. R. Karim *et al.*, "Interpreting black-box machine learning models for high-dimensional datasets," in *Proceedings of the 2023 IEEE 10th International Conference on Data Science and Advanced Analytics (DSAA)*, 2023, pp. 1–10. [Online]. Available: <https://doi.org/10.1109/DSAA60987.2023.10302562>
- [39] I. A. V. Ikechukwu and S. Murali, "xAI: An explainable AI model for the diagnosis of COPD from CXR images," in *Proceedings of the 2023 IEEE 2nd International Conference on Data, Decision and Systems (ICDDS)*, Mangaluru, India, 2023, pp. 1–6. [Online]. Available: <https://doi.org/10.1109/ICDDS59137.2023.10434619>
- [40] A. Salih, I. Boscolo Galazzo, P. Gkontra, A. M. Lee, K. Lekadir, Z. Raisi-Estabragh, and S. E. Petersen, "Explainable artificial intelligence and cardiac imaging: Toward more interpretable models," *Circulation: Cardiovascular Imaging*, vol. 16, 2023. [Online]. Available: <https://doi.org/10.1161/CIRCIMAGING.122.014519>
- [41] A. M. Conard, A. DenAdel, and L. Crawford, "A spectrum of explainable and interpretable machine learning approaches for genomic studies," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 15, p. e1617, 2023. [Online]. Available: <https://doi.org/10.1002/wics.1617>
- [42] R.-K. Sheu and M. S. Pardeshi, "A survey on medical explainable AI (XAI): Recent progress, explainability approach, human interaction and scoring system," *Sensors*, vol. 22, no. 22, p. 8068, 2022. [Online]. Available: <https://doi.org/10.3390/s22208068>
- [43] R. El Shawi and M. H. Al-Mallah, "Interpretable local concept-based explanation with human feedback to predict all-cause mortality," *Journal of Artificial Intelligence Research*, vol. 75, 2022. [Online]. Available: <https://doi.org/10.1613/jair.1.14019>
- [44] J. El Zini and M. Awad, "On the explainability of natural language processing deep models," *ACM Computing Surveys*, vol. 55, no. 5, pp. Article 103, 1–31, 2023. [Online]. Available: <https://doi.org/10.1145/3529755>
- [45] S. A. Dubey and A. A. Pandit, "A comprehensive review and application of interpretable deep learning model for ADR prediction," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 13, no. 9, 2022. [Online]. Available: <https://doi.org/10.14569/IJACSA.2022.0130924>

Citation: M.C. Pezzini and C. Pons. *Explainable Artificial Intelligence: Analysis of Methodologies and Applications*. Journal of Computer Science & Technology, vol. 25, no. 2, pp. 75–86, 2025.

DOI: 10.24215/16666038.25.e07.

Received: March 26, 2025 **Accepted:** October 14, 2025.

Copyright: This article is distributed under the terms of the Creative Commons License CC-BY-NC-SA.