# Explainable AI (XAI) Methods: Interpretability, Trust, and Applications in Critical Systems: A Systematic Literature Review.

WASIU OLATUNDE OLADAPO[1], ISMAIL OLANIYI MURAINA[1,] MOSES ADEOLU AGOI[1], SOLOMON ONEN ABAM[2], BASHIR OYENIRAN AYINDE[3]

Information and Communication Technology (ICT)
[1]Lagos State University of Education, Lagos.
Km 30 Badagry Express Way Oto/Ijanikin, P.M.B 007 Festac Town, Lagos State.
NIGERIA.
Computer Science Education
[2]Federal College of Education, Technical, Isu, Ebonyi State
NIGERIA.
Computer Science Education
[3]Lagos State University Africa Center of Excellence for Innovative and Transformative Stem Education, Ojo, Lagos,
NIGERIA.

*Abstract:* - The systematic literature review study examines more recent advances in Explainable Artificial Intelligence (XAI) under the umbrellas of interpretability, trust, and application in critical systems. The research is a study synthesis, which analyses the findings of 18 peer-reviewed articles published between 2020 and 2025, providing a synopsis of the XAI frameworks and methods and their domain-specific applications. The most promising XAI tools like LIME, SHAP, counterfactual explanations, and model-agnostic methods are analyzed in the different fields of various applications: healthcare, cybersecurity, finance, industrial control systems, and autonomous vehicles. The review points out the conflict between accuracy and interpretability of models, and a possible absence of a standard metric used to measure the quality of an explanation procedure. It also highlights the necessity of user and context-based explanations in assisting high-stakes environments in making decisions. Ethical aspects, humanity, and security as well as the industry-related issues on trust and safety are critically evaluated. Both in the number and in the magnitude, there has been an increased traction of XAI researches, especially as of 2024 to 2025, and the future direction of interest suggests the necessity of future research into scalable XAI techniques, evaluation venues, and the involvement of large language models to provide natural language explanations. The review helps to promote trustworthy AI by highlighting the research gaps, summarizing the trends and suggesting best practices on the deployment of explainable systems as applied to mission-critical tasks.

*Key-Words:* - Explainable AI (XAI), Interpretable AI, AI explainability, Interpretability, Transparency, Explanation, Understandable, Trust, , Critical systems and Application.

Received: June 28, 2025. Revised: August 4, 2025. Accepted: August 27, 2025. Published: October 22, 2025.

## 1 Introduction

The goal of explainable AI (XAI) is to increase the visibility and interpretability of black box AI systems, particularly in areas where the lives of humans, legal fate, financial stability, or community safety might be at risk. Core XAI approaches involve interpretable model construction (e.g. decision trees, rule-based systems) and post hoc explanation techniques such as LIME, SHAP and counterfactual explanations that enable black box models to be better explained to their end users. The LIME method can be described as the approximation of local behaviour of the complicated models to simpler ones fitted around the points of interest. SHAP uses cooperative game theoretic Shapley values to assign feature importance to offer both the global and local interpretability. Counterfactual reasons also provide what-would-have-been transformations essential in fields where there must be equity or law responsibility (e.g. lending choices, criminal justice risk-ranking) [1]. In clinical decision support, [1] propose that

transparency is needed to win the trust of clinicians, that explainability needs to be supplemented by external validation and reporting of data quality and interaction with stakeholders. [2] Integrate ML explanation models into the actual clinical process and discover a non-trivial impact: explanations do decrease automation bias and aid in ambiguous cases, as well as support junior clinicians but can also introduce confirmation bias or cognitive overhead. [3] Propose a design framework of Trustworthy AI in health care, which extends beyond explainability into traceability, communications, and explanation robustness including supporting rich set requirements like multiple customizable levels of explanation and visual techniques. The study by [4] indicates its gaps and makes suggestions of the future work such as user-centred evaluation and lifecycle aware integration of explainability during the design, development, and deployment life stages. A 2022 survey also talks more about the particular schema to compare XAI methods through the respective levels of consistency, complexity, and measures of causal reasoning and proposes BRB systems in the time-series explainability in important areas. The article written by [5] offers a breakdown of autonomous cyber-defence systems. It captures how the users like transparency but they undergo barriers due to complexity and training. The study implies comparing the XAI methods and outlining user education to build trust. However, when describing another study regarding IIoT in 2022, the introduction of a model agnostic statistical explainability framework with industrial security (IIoT) called TRUST XAI has been introduced due to the achievement of the explanation success of ~98% and can be efficacious and explainable as compared to LIME. XAI is highly needed in the healthcare setting where physicians require explanations before adopting any recommendations (e.g. detection of cancer based on CT scans). [1] Offer advice on XAI design used in clinical decision support regulators, making distinctions between explained modelling approaches and post hoc descriptions, the necessity of validation and meaningful measures that lead to lucidity and fidelity.

XAI is used in autonomous cars and industrial control to enable safety, situational awareness and regulatory transparency. The interpretable rules and visualizations provided by prototype-based models allow making sense of decisions in unusual cases or edge cases and understanding how to act when not following an existing pattern [6]. XAI is more frequently involved in the explanation of models of air quality, water pollution, and climate events prediction. Other approaches, such as SHAP when applied along with environmental models, facilitate the interpretation of changing factor effects on a phenomenon such as a pollution outbreak, or a temperature deviation [7].

## 1.1 Limitations of the existing knowledge

### 1.1.1 Scalability Model & Complexity Existing

XAI methods have presented serious difficulty to the modern AI systems, including deep neural networks, transformers, and ensemble models. Such methods as the LIME and SHAP are not good at scaling to high dimensional data, and at capturing global behaviour of models, they are either computationally demanding or unreliable. They are usually unable to trace through the nonlinear, complex dependencies limiting trustworthiness in problem critical fields (e.g. medical imaging, and autonomous systems) [8].

### 1.1.2 Accuracy Vs. Interpretability Trade Off

It is still not certain that there exists a fool proof method that would achieve high interpretability and predictive accuracy. Simple models (such as intrinsically interpretable models, e.g. decision trees, or GAMs) tend to fail on vital tasks when compared to complex ones, and post hoc explainability techniques can simplify results too far. Such trade-off is of particular concern in areas where transparency and performance are critical, e.g. healthcare, finance, or autonomous systems [8].

### 1.1.3 The absence of a Standardized Evaluation Metric

The discipline is thus lacking strong evaluation criteria regarding the quality of explanations, which are standardized. Numerous studies are based on either subjective or proxy tasks where the performance of the AI is compared to the guesses made by users as opposed to the actual performance of human plus AI in domain-relevant decision tasks. Also there can be no benchmarking between methods without known ground truth on explanations, and even the results will be difficult to compare and generalize.

### 1.1.4 Scanty Empirical evidence of contributions to Decision Making

Based on experimental research, there is no universal guarantee that any form of explanation can serve to improve human decision-making. Under controlled experiments, users showed better results with guidance by AI but not with the

provision of explanation which contributed little to the increment in trust and decision accuracy with regard to the basic performance of the AI. This is a challenge to the supposition that explainability can have a direct effect as a means of producing superior results in life-and-death tasks.

#### 1.1.5 Explainability vs. Trust ambiguities

Philosophical and empirical studies have shown that sometimes explainability is not the necessity in creating trust, particularly, forms of trust which is suitable to machine based systems. Certain types of trust can be based on reliability, validation and/or feedback loops, and not necessarily because of pure explanation. Practically, users can comprehend an AI decision and at the same time doubt the system especially when human judgment is desired, especially in high stakes situations.

#### 1.1.6 Human Factors & Cognitive Variability

The existing XAI do not necessarily provide explanations made at the discretion of AI researchers instead of the end users or the domain experts. These explanations can be too technical or abstract to make comprehension slow and, thus, unable to aid decision-making. Context-sensitive, intelligent explanation rules have not been developed fully, and thus the user can be misperceived or become biased or overwhelmed.

#### 1.1.7 Security, Privacy and Ethical risks

The exposure of proprietary logic, or artifacts used in training data and/or vulnerabilities, may result unintentionally when making AI models more explainable. Evil users might reverse-engineer systems, or users might "game" explanations, which destroys fairness or optimality (observed in hiring or predictive policing) examples. Descriptions may also leak sensitive personal information or model internals, potentially creating compliance and privacy issues.

#### 1.1.8 Sector-Specific Safety versus Trust-Focus

A large number of reviews and papers focus on XAI as a method of increasing confidence in applications especially in the healthcare and cyber security domains but do little to address concerns of safety or correctness in critical decision systems. Sensitivity checking, fault-detection and integrity verification are not well represented in theory and empirical research. The holistic cross-sector analysis between the healthcare, critical infrastructure and industrial automation is not common.

### 1.2 Research Questions

1. When applied to non-technical end-users of AI decision-making systems, do post-hoc XAI tools like SHAP or LIME in an otherwise opaque black-box reduce calibration of trust/ credibility, interpretability, and acceptance of AI-driven decisions as compared to black-boxes without explanations?

2. When applied to a high-stakes application such as healthcare, finance, or criminal justice, in an AI application, will the domain-specific and user-customized XAI techniques perform better than a generic explanation model, in terms of user understanding and cognitive load?

3. Between the model-agnostic and model-specific/intrinsic interpretable models, do model-agnostic XAI tools offer improved trade-offs to explainability versus predictive accuracy in practice among AI developers and data scientists working with black-box models?

4. Is the use of real-time, adversarial-robust XAI deployments, in comparison with classical, non-real-time feature explanation techniques, beneficial to the interpretability of threat detections in systems deployed in cybersecurity without reducing the robustness of those systems?

5. When AI systems with natural language generation explanations are explainable, or contain large language models to explain the outputs, does multimodal vs. text-only explanation increase user satisfaction, and comprehension?

## 2 Methodology

The aim of this systematic literature review (SLR) was to investigate and examine the application of explainable artificial intelligence (XAI) method to interpretability, trust, and applications in critical system. Databases (Google Scholar, Semantic Scholar, ResearchGate ScienceDirect, IEEE Xplore, Wiley Online Library, Crossref, COinS, ojs.boulibrary.com) have been chosen as databases to cover the relevant explainable artificial intelligence literature fully and have been used as the exclusive database sources in this review. Google Scholar, Semantic Scholar and Research

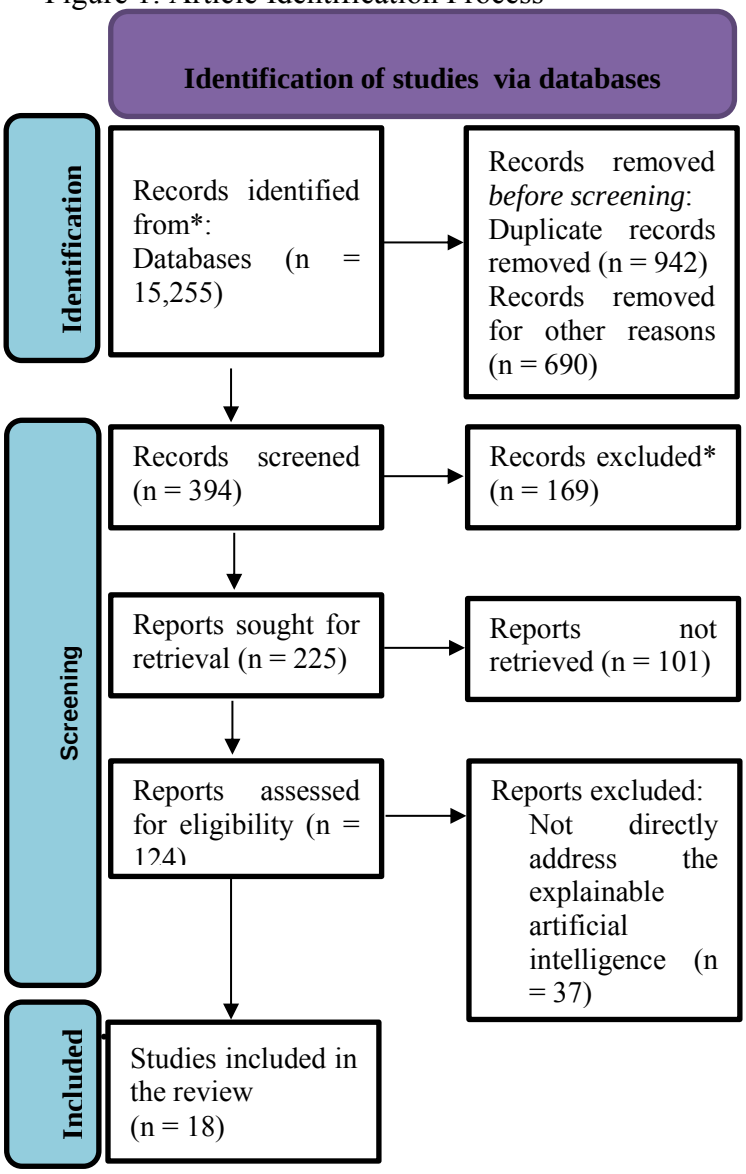
Gate were high-profile and rich databases that allowed access to peer-reviewed journals, conferences papers, and book chapters in various fields, which made them very suitable to a narrow search of the field of interest [46]. The search query was run (explainable AI" OR XAI OR "interpretable AI" OR "AI explainability") AND (interpretability OR transparency OR explanation OR explainability OR understandable) AND (trust OR confidence OR accountability OR reliability OR "human-AI trust") AND (methods OR techniques OR frameworks OR algorithms OR models) that returned the initial set of 15,255 documents.

These documents had varying areas, so to narrow the search down and to cue on the developments made in the area in recent years the publication year was restricted to a span of between 2020 and 2025. This filter narrowed the document set to 942. To further reduce the scope, the search scope was reduced to a subject area to computer science further reducing the number of documents to 690. Since the review was restricted to primary research, the types of documents/literatures considered were only articles; limiting the number of documents to 394. In case only the finalized research should be put into consideration another filter was used that confined the search to the stage of final publication leaving behind 225 documents. Finally, the type of source was limited to journal papers and the language was to English, which gave a total of 124 articles.

In order to search more specifically studies that could be used to explain explainable artificial intelligence and areas of application in systems with critical functions, certain keywords were used in the search process, as follows: “explainable AI,” “interpretable AI,” “AI explainability,”“interpretability,” “transparency”, “explanation”, “explainability”, “understandable”, “trust”, “confidence”, “accountability” and “reliability,” as well as a selection of words used in applicative fields of various disciplines, such as: “techniques”, “methods”, “techniques”, “ The reason as to why these keywords were used is that they guaranteed that the chosen studies directly exposed the explainable artificial intelligence in the computer science field. A review of the 124 documents then produced an Excel sheet that was used to screen and analyze documents further. Since the 124 documents that were exported out of the databases to an Excel sheet a duplicate removal process was the next to be carried out. This was necessary to make sure that repeated articles in the first search were given and would not be used in subsequent analysis. After removing the duplicates, the abstracts of the remaining documents were thoroughly read and appraised in an attempt to determine their relevance to the research questions of the review. Such abstract screening permitted the removal of the studies that were not directly related to the explainability of artificial intelligence in the fields of computer science. This way, the screening process resulted in the narrowed and final list of 18 documents which, in their turn, were chosen to be reviewed with full-text and thoroughly analysed.

**The below PRISMA model illustrates the systematic procedure followed in the identification, screening, and selection of articles.**

Figure 1: Article Identification Process



# 3 Findings and Discussion

Due to its detailed review and discussion of 18 selected research papers, the findings and discussion presented in this systematic literature review focused on five (5) topics including: non-technical end-users of AI driven decision systems, AI systems in high stakes domains including healthcare, finance, or criminal justice, do domain-specific among AI developers and data scientists who work with black-boxes models, AI based systems implemented in cybersecurity positions, and explainable AI systems written through large language models to dynamically generate natural language explanations. Each section consists of the results of the various studies available, demonstrating new trends, significant breakthroughs, and unresolved issues in the realm of the artificial intelligence domain.

## 3.1 Trend of articles on explainable AI (XAI) methods: interpretability, trust and applications to critical systems publication

The analysis of (Table 1) shows that the focal point of the research lies with AI (all papers) with a total of 18 papers, unique authors are 43, and highly emphasized topics are on explainability (13 papers), trust (14 papers), explainable artificial intelligence (14 paper) all the paper review on artificial intelligence (17 paper) and most common publication year is 2024 (with 9 papers). The publication pattern is that there is a rising trend, with the year 2025 being most prolific. The mean rate of collaborations is 2.39 authors per paper, which was rather moderate.

**Table 1: Research Analysis**

| Research_Analysis | |
|---|---|
| **Metric** | **Value** |
| Total_Papers | 18 |
| Unique_Autho | 43 |
| Average_Auth | 2.39 |
| Most_Commo | 2024 |
| Most_Commo | 9 |
| Trust_Mention | 14 |
| Explainable_M | 13 |
| AI_Mentions | 17 |
| XAI_Mentions | 14 |

The graph below in figure 2 indicates that 2025 is a break out year implying that this is an upcoming and developing field of research. The growing step suggests that more academic focus and probably, the significance of explainable AI are rising in different applications. The research project begins in a humble way in 2024 with a small number of papers whereas in year 2025 it expands massively indicating that the field is rapidly growing. The trend points to an increasing interest and momentum in the AI explainability research. Although a steady growth is observed when cumulative line is used, having 18 total publications by 2025.

The chart shows a clear growth pattern in AI explainability research over time:

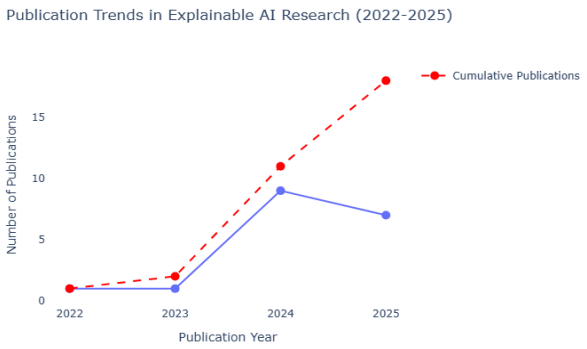



**Fig. 2:** Publication Distribution per Year

The list of 18 research publications is between 2022 and 2025 and broadly describes topics of Explainable Artificial Intelligence (XAI) and its development in different areas.

### 3.1.1 Most important Research Themes and Results:

Fundamental XAI Frameworks and Methodologies: A number of works are aimed at the theoretic framework and taxonomies of XAI. [8] Suggests a unified taxonomy of explanations approaches that includes the cognitive theory and [9] propose a modular framework of reliable robotics Systems. Such establishment works stress out the significance of conventionalized procedures on equity, responsibility, and ethical implementation of AI.

Systematic Reviews and Comparative Analyses: Various systematic reviews of the literature indicate important information about the situation of XAI research. In a systematic review, [10] determine the gaps in evaluation criteria and user-centered investigations of XAI techniques to support human trust. [11] Provide a critical analysis of XAI applications that channel the further direction of research on them. [12] Provides a

systematic review of XAI methods with the goal to set best practices concerning implementation practice.

Domain-Specific Applications: The study embraces various XAI applications in industries:

1. Healthcare and Finance: [13] creates models of trust assessment of AI in crucial fields, focusing on the human-AI collaboration.
2. Cyber-Physical Systems: [14] investigate multisensory description of CPS industrial, and pay attention to ethics and safety.
3. Renewable Energy: [15] investigate trustworthy AI methodologies for energy systems decision-making Cybersecurity. [16] Scrutinize XAI methods of threat analysis and transparency in security.
4. Emergency Management: [17] discusses the AI transparency and building trust in a crisis response system.

Technical Innovations and Tools: A number of research papers are devoted to a particular technique and tool of XAI: [18] talks about feature importance and surrogate models that employ LIME and SHAP.[19] examines transparent methods of interpretable deep learning on decision support.

Ethical and Trust Considerations: One aspect present throughout the studies is the ethical consequence of AI transparency. [20] Run experiments on the aspect of trust in AI systems, and [14] focus on the ethical side of multisensory explanation. According to these studies, it became obvious that it is essential to balance the technological progress with human control to make the genuinely responsible AI implementation possible.

### 3.1.2 Possible Benefits at the Multiple Landscapes:

1. Better decision-making and accountability in autonomous systems, healthcare and finance.
2. Increase in confidence and detecting errors of users by transparent AI models.
3. Ethical deployment of AI with standardized fairness protocols.
4. Efficient use of resources on the agricultural and energy fronts.
5. More AI use and confidence in infrastructure provision, emergency response.

The given body of the research contributes to the overall XAI field, considering both theoretical innovations and practical aspects of transparency, trust, and ethical concerns in the AI system implementation into various spheres.

### 3.1.3 Non-technical end-users of decision systems driven by Artificial Intelligence

In their analysis, [21] discuss the meaning of creating explanations in non-technical audiences, highlighting that the focus of much XAI is on expert-level users and claiming that instead to enable explanation modalities with laypeople, approaches must be based on field implementations. They give rules and a case study in a controlled field where domain professionals and not AI experts use them mostly. The paper by [22] discusses the end user expectations concerning explanations in recommendation systems. Using focus groups, they discover that non-technical users have a preference on a tailored, on-demand and privacy-sensitive explanation with trust being more strongly dependent on comprehensibility than technical detail. [23] are able to provide a systematic review of user trust in the AI-enabled systems in the HCI (Human Computer Interaction) lens. They emphasize how trust is highly reliant on socio ethical issues, user characteristics and design, and that trust should be established non-technically using users in the development process. [24] Directly includes the concept of human centric usability when considering AI driven clinical decision support systems (CDSS) such as transparency, interpretability, trust, and an iterative design with the end user being clinical professionals rather than experts in AI. In a systematic review, [25] discuss the system of strategic decision making in enterprise platforms. They observe that the results of the decisions and trust can be affected by factors such as user expertise, the organizational chain of command and culture when deployed in non-technical personnel working with AI advisory support tools. Algorithm aversion was observed in the literature of how non-technical users do not believe algorithm suggestions, especially when they have an error. Human in the loop mechanisms, transparency, personalization and the ability to enable users to exercise control over the outputs by algorithms, can address this aversion. The research in automation bias presents a corresponding problem; there is a tendency of users in default trusting AI at the expense of the internal contradictory information. Interface and training design can assist non-

technical users to be on alert and make good decisions.

### 3.1.4. Domain-specific AI Do AI systems that work in high-stakes fields (healthcare, finance, or criminal justice, etc.) perform domain-specific

[26] Provide a detailed literature review of the risks related to the use of AI in healthcare and identify the risks of clinical-data, technical, and socio ethical risks. They discuss 39 papers and pose the system of risk mitigation in high-stakes clinical systems. The article by [27] provides a survey of problems of fairness and bias in healthcare, finance, and criminal justice, emphasizing how making datasets, algorithms, and other technologies biased perpetuates inequality and suggesting addressing them by employing an interdisciplinary mitigation approach. Deeper inspection highlights the issue of the "black box" of healthcare AI influencing the trust and regulatory compliance and the necessity to consider transparency, explainability, and the ethic of not doing harm. [28] Article cites the so-called accountability dilemma associated with the AI-driven systems in the finance industry making high-stakes decisions such as credit scoring or trading but remaining obscure. Regulators and financial heads are promoting the move to a more explainable and interpretable system, black-box for glass-box through either XAI, blockchain, or more stringent governance principles. [29] Provide a systematic review of 37 articles involving AI in criminal justice management to discuss the areas of predictive policing, risk assessment, surveillance, as well as present various ethical, fairness, transparency and regulatory issues. A survey of 66 scholarly works in an article conducted by [30] revealed that the most common of them is the application of AI in predicting crime and making legal decision support. They emphasize the two-fold role of AI in efficiency gains and legal/ethical risk. [31] Address bias in computer vision on criminal justice e.g. facial recognition systems applied by the police and suggest mitigation options to balance fairness in the development and implementation of the systems. Survey-based experiments among legal and law enforcement professionals indicate that [32] found skepticism toward AI, however, large proportions reporting willingness to follow algorithmic advisory in probation, sentencing, and warrants, often to a greater extent than peer judgments.

**Themes Criminal justice**

Perhaps the most widespread example of effective use, however, is facial recognition, surveillance, and COMPAS risk tools.

Areas of concern: bias in algorithms and lack of transparency, questions of unfairness, and Practitioner trends: perceived distrust and increased practical dependence on the outputs of the artificial intelligence system.

**Comparisons across Domains and Syntheses**

[33] Explore uses of Large Language Models (LLMs) such as GPT 4 in finance, health systems, and law, with common themes of regulatory limits, legal-financial stakes of accuracy, equity, and need of domain-specific interpretability. [27] Outlines challenges of fairness in different areas and highlights typical causes of bias in training data, algorithm construction, human interaction and appeals to domain-sensitive mitigation schemes.

### 3.1.5 AI Developers and Data Scientists who use Black-Box models

The survey by [34] explores the question of how developers and data scientists interpret the principles of deep neural networks (DNNs) the most popular black-box models. It addresses algorithms used in interpretation, metrics of evaluation, and the make-up of trust in interpreting outputs in models. The authors emphasize the accuracy-explainability trade-off that is in the control of data practitioners and the choice of interpretation libraries. [35] It was specially designed to help data scientists by providing a report on a taxonomy and decision tree to select explanation methods (e.g. SHAP, LIME). It discusses developer reasons to use explainability tools in consideration of performance, model type and stakeholder interests. The review overviews the way, in which the various model developers have interrogated black-box models through XAI, with interests in visualization, feature ranking and local or global explanations. To interpretability, data scientists perceive interpretability as an absolute when only precision models pose a threat to regulated or high-stakes systems. There is a

review of explainability frameworks which are deployed by developers to help people understand (e.g. interactive visual tools, visualization dashboards (e.g. GANViz, DGMTracker). It tells about the way developers visualize training phenomena and how such tools help them to debug and reduce bias in black-box designs.

This paper provides a road map to reliability and interpretability of models to the practitioner of machine learning in the black-box setting. It describes tradeoff balances between model accuracy and transparency, and how approaches such as counterfactual explanations or post hoc attribution strategies are being used, either leaving the underlying model unaffected. There is a trade-off between preserving high levels of predictive performance and satisfying stake URL demands of auditability and trust as technical engineers are minded to gain. The explored framework, which is presented by [36], offers a taxonomy that can be used to evaluate explainable systems in a systematic manner along such dimensions as functional performance, usability, safety, and validation. The "Explainability Fact Sheets" assist the practitioners to write down and assess the characteristics, usages, and constraints of the XAI methods. The Model Usability Evaluation Framework (MUsE) identified by [37] is supposed to evaluate the efficiency of explainable AI tools with a user experience (UX) of an engineer, in particular, to examine their interaction with a model-agnostic explainer such as LIME. It draws attention to post-hoc explainability UX trade-offs without model change. Research goes on to point out that developers do not consider the ethics implication: although in some cases the developer applies black box models, the documentation seldom has any indication in mitigating risk. In the technical documentation transparency is often loosely defined. This concern is augmented when operating in teams that use proprietary frameworks as the developers themselves might not be able to follow the decision reasoning since models are not well visible [38]. A cognitive-computation survey highlights that human control is an increasingly employed feature by developers in black-box systems. The combination of the feedback loops and interactive dashboards is expected to allow practitioners to increase transparency, identify bias at the first stage, and get the improvement in performance on the long-term basis. Explainability tools are not only seen as a requirement by developers to their end users in-house validation, bias reduction and enhancement also require explainability tools.

### 3.1.6 AI driven systems that are implemented in cyber defense situations

[39] Provides a systemic literature review with sorts of ML, DL, and NLP-based applications in cybersecurity in threat detection, anomaly assessment, or incident response. The paper presents the reflection on the ethical and financial issues raised by the implementation of AI in multifaceted security settings. In their review entitled Explainable AI (XAI) in cybersecurity. [40] Outline the current XAI approaches adopted in an attempt to enhance the following such transparency in operator-facing systems, and what cybersecurity areas are already using XAI techniques. [41] Explores many deep learning models (CNN, RNN, DBN, LSTM, auto encoders, hybrid model) used in network intrusion detection, measurement of the detection performance, resource consumption, and scalability of the models on benchmark datasets. A systematic review of 58 IDS studies based on ML and DL is given by [42]. They compare classical methods (SVM, Random Forest, and KNN) with deep architectures and compare them on KDD cup, NSL KDD, UNSW NB15 and Kyoto data. They got an RF ~99.5 accuracy, whereas CNNs and MLPs were fast in precision even though they conducted prolonged training. [43] Analyses the intrusion detection technique based on deep learning in comparison to CNN, RNN, auto encoder, and hybrid approaches. It tests these in real time performance and resource consumption as well as compliance in complex network data environments. An example of stacked non symmetric deep auto-encoder with SVM described by [44] provides 99.65 accuracy, 99.99 precision, and low false alarms on the KDD 99 dataset indicating how hybrid DL + classical ML can have even better detection rates than pure ML approaches can achieve. The review identifies issues like imbalance in dataset, adversarial susceptibility, absence of model explanation and combination with edge/IoT conditions. [45] Propose federated learning of IDS, which requires privacy-saving training at the edge sources and centralization of only model parameters- appropriate to sensitive datacenters where there is distributed work. [40] Stress that when developing AI detection systems, it is possible to use opaque models and suggest that more XAI tools should be utilized so that cybersecurity operators can comprehend and feel confident in the decisions that are made

automatically. Malicious machine learning has significant security hazards: adversarial and malicious machine learning threats can be used to subvert or interfere with AI based defenses. Model design should take into consideration adversarial robustness because of security aware approaches. According to [11], the organization advises a combination of using AI on the top of the conventional security-related analytics and keeping human AI cooperation. It cautions against vulnerabilities bias, privacy risks, prompt-injection and emphasises the synergy between AI instruments and human analysts.

#### 3.1.7 Explainable AI systems using large language models for natural language generation of explanations.

This survey offers taxonomy of explainability methods to Transformer-based LLMs. It differentiates local and global explanations, including the paradigms of fine-tuning, and paradigms of prompting. The authors evaluate the main issues related to such aspects as scalability, faithfulness, and user-centered measures of evaluation. This is among the limited detailed reviews that both integrated explanation generation in the context of LLM. [47] Presents a wider scope of the XAI approaches that are developing into LLM-translated explanations. It places LLMs as the subject of explanation as well as an explanation-generating tool. This paper reviews the extent to which LLMs can automatically convert model outputs (e.g., numeric attribution scores) into narrative explanations that are semantically rich (e.g., words that are relevant and explainable to a target audience rather than unstructured raw text). [48] Elaborate on a two-part system based on LLM, whose components consist of a Narrator (which converts SHAP or other attribution-based explanations into coherent narratives) and Grader (which assesses the fluency, completeness, and accuracy). Experiments also show that the LLM-generated explanations are scored highly by humans, and are much better at improving end-user understanding in the case of non-technical stakeholders. [49] Examines the potential to use traditional model outputs (feature importance, counterfactuals, scores) to fully feed risk assessment systems to LLMs and generate explanatory results. The author addresses design specifications, the assessment of the narrative quality, and area-on-pointed elaboration strategies. [50] Discusses the ways a more transparent LLM (e.g., GPT, PaLM, LLaMA) can be achieved. It discusses techniques like attention visualization, Integrated Gradients, causal inference and neuro-symbolic hybrid methods. Although it does not necessarily focus on natural-language explanations, the paper gives key insight into the trade-offs between explainability and model performance that are especially pertinent when LLMs are supposed to fulfill dual purposes [50] surveys LLM-specific XAI methods: feature attribution over tokens, attention analysis, causal reasoning, and counterfactual generation. The paper identifies three major challenges that include scaling XAI methods to billion-parameter models, narrative faithfulness of explanations, and misleading or hallucinated content in actual explanations. According to the review given by [49], these systems are interactive XAI workflows that use LLMs. As an example, Chat conversation enables the operators to pose follow-up questions and a natural language explanation composed of SHAP or LIME values. Such platforms (e.g., HuntGPT) apply in cybersecurity scenarios, where human-in-loop interpretation and intervention are applicable.

#### 3.1.8 Assessment and Human-Based Metrics

The quality of explanations is also often judged on accuracy, completeness, fluency and conciseness, as evidenced by the Grader component of the Explingo system and by user studies which found users significantly preferred LLM generated explanation results to raw attack attribution output. What is also raised in the literature is the issue of faithfulness, the explanation generated by LLM can be fluent but wrong or misleading. De-escalating hallucinations and being consistent with the source attestation data is a primary open subject.

## 4 Conceptual State of Art on XAI research

The 18 screened articles, together, offer an overview of the panorama of the present and future of XAI, outlining a number of important trends and research priorities:

### 4.1 Methodological Diversity; Framework Development

1. Theoretical Foundations: Research articles such as [8], and [9] are concerned with coming up with sound theoretical frameworks and designs of XAI. These papers highlight the necessity to have a standard taxonomies and modular design that can be incorporated between the various AI applications.

2. Systematic Reviews: Multiple large bibliographic reviews [10], [11] and [12] demonstrate the heterogeneity of the existing XAI approaches and outline the essential research areas that remain unexplored, including standardized metrics of evaluation and more user-centered research.

### 4.2 Application-Specific Insights

1. Merchandise and banking: trust analysis models and understandable deep learning approaches [13]. [19] Show transparency is critical to the high-stakes decision-making context.
2. Industrial and Cyber-Physical Systems: [15] and [14] demonstrate that a multisensory explanation approach and trustworthy AI techniques are essential to security, fairness, and reliability in the context of complex industrial systems.
3. Agriculture and Energy: [50] and [15] portray the XAI tool applied to sustainability-related sectors and decision-making processes about the efficient use of resources.
4. Imperfect solutions: [16] and [17] note that operating with the assumptions that current AI systems lack transparency; it is possible to create simple systems that build trust in both critical infrastructure and crisis response situations.

### 4.3 Technical Solutions and Equipment

1. Surrogate Models and Feature Importance: [18] and [19] point to the quality of such methods as LIME and SHAP to make complex models easier to interpret. Also, deals with the problem of ethical issues of AI personalization, suggesting XAI methods of identifying and minimizing bias in AI.
2. Human-Centered Design: [9] propose the hybrid models that would provide a compromise between performance and interpretability to ensure that the actions of the AI systems are understandable to humans.

### 4.4 Ethical and Trust matters

**1.** Trust Factors: [20] present the experimental data on what can influence human trust in AI, and it is a crucial aspect to consider the transparency, reliability of the product and respect ethics.

**2.** Ethical Deployment: [14] emphasize that there must be an ethical framework which will make AI systems not only efficient but also fair and accountable.

**3.** Transparency Strategies: [17] present the viable strategies of establishing trustworthiness and accountability in the AI systems that operate in critical areas.

### 4.5 Addressed Research Gaps and Future Research

1. Standardization Requirements: Multiple reviews outline the necessity of standardized measures of success to enable the methods effectiveness to be compared across techniques.
2. User-Centric Evaluations: The necessity of additional research that covers the interaction of a user with and the comprehension of AI explanations becomes apparent.
3. Long-Term Effects: There are limited studies that have investigated the impacts of XAI with time on user trust and decision.
4. Cross-Domain Adaptability: There is more work required in the field of how the XAI methods might be applied across various industries and applications.

## Synthesis

The XAI reviewed papers altogether prove that effective XAI can:

1. Improvement of Decision-Making: XAI allows users to make more and more confident decisions about their data.
2. Develop Responsibility: Clear AI systems form more accountable systems, easier to audit and regulate.
3. Enhance Teamwork: The trustworthy AI can promote the work of humans and AI to collaborate in areas where greater work is required.
4. Ethical AI: By means of the XAI methods, it is possible to identify and reduce biases to provide more ethical results.
5. Amplify Adoption: Understandable and Explainable AI systems would tend to attract more users and stakeholders

**Table.2: Summary Table of Reviewed Papers**

| S/N | Authors | Year of Publication | Technology Used | Key Features | Potential Benefits |
|---|---|---|---|---|---|
| 1 | Barnes, Emily; Hutson, | 2024 | Deep Neural Networks (DNN | Architectural design choices impact interpr | Improved decision-making, |

| | | | | | |
|---|---|---|---|---|---|
| | James | | s) | etability; LIME and SHAP tools used | accountability, regulatory compliance in healthcare, finance, autonomous driving |
| 3 | Arunraju Chinnaraju | 2025 | Theoretical XAI Framework | Taxonomy of explanation methods; cognitive theory integration | Standardized protocols for fairness, accountability in AI deployment |
| 4 | Adebayo, Abiodun Sunday; Ajayi, Olanrewaju Oluwaseun; Chukwurah, Naomi | 2024 | XAI in Robotics | Modular architecture, hybrid models, human-centered design | Trustworthy, reliable robotic systems with balanced performance and transparency |
| 5 | Kanagarla, Krishna Prasanth Brahmaji | 2024 | XAI in Data Analytics | Feature importance, surrogate models, LIME, SHAP | Transparency in complex ML models, better stakeholder understanding |
| 6 | Guillaume Jean | 2024 | Trust Evaluation Models | Feature importance analysis, user feedback mechanisms | Improved human-AI collaboration in healthcare, finance, autonomous systems |
| 7 | Amber Hoenig; Roy, Kaushik; Acquaah, Yaa Takyiwaa; Yi, Sun; Desai, Salil S. | 2024 | XAI for Cyber-Physical Systems | Multisensory explanations, ethical considerations | Enhanced control, fairness, safety in industrial CPS |
| 8 | de Brito Duarte, Regina; Corr | 2023 | AI Trust Studies | Experimental analysis of trust factors in AI | Understanding conditions for effecti |

| | | | | | |
|---|---|---|---|---|---|
| | eia, Filipa; Arriaga, Patrícia; Paiva, Ana | | | systems | ve AI trust building |
| 9 | Amith Kumar Reddy | 2024 | Interpretable Deep Learning | SHAP, LIME, model-specific interpretability | Transparent AI decision support systems across industries |
| 10 | Kalasampath, Khushi; Spoorthi, K. N.; Sajeev, Sreeparvathy; Kuppa, Sahil Sarma; Ajay, Kavya; Maruthamuthu, | 2025 | XAI Applications Review | Systematic literature review of XAI techniques | Guidance for future XAI research and applications |
| | Angulakshmi | | | | |
| 11 | Para, Raghu K. | 2024 | XAI for Bias Mitigation | Explainability techniques for bias detection | Ethical hyper-personalization strategies with reduced bias |
| 12 | Mohyuddin, Ghulam; Khan, Muhammad Adnan; Haseeb, Abdul; Mahpara, Shahzadi; Waseem, Muhammad; Saleh, Ahmed Mohammed | 2024 | ML in Precision Farming | Comparative analysis of ML algorithms | Optimized crop management and resource efficiency |
| 13 | Ersöz, Betül; | 2022 | XAI in Renewabl | Trustworthy AI metho | Enhanced decisi |

| | | | | | |
|---|---|---|---|---|---|
| | Sağıroğlu, Şeref; Bülbül, Halil | | e Energy | dologies | on-making in energy systems |
| 14 | Sarela, Mirka; Podgorelec, Vili | 2024 | XAI Applications Review | Comprehensive literature analysis | Identification of research gaps and future directions |
| 15 | Jonqobilov, Mirjalol | 2025 | AI Trust Frameworks | Strategies for transparency and trustworthiness | Improved AI adoption in critical sectors |
| 16 | David Alexander; Maria Aaron | 2025 | XAI in Cybersecurity | Explainability techniques for threat analysis | Enhanced transparency and trust in security systems |
| 17 | Ayokunle Micheal Akinsiku | 2025 | XAI Techniques Review | Systematic review of XAI methods | Best practices for XAI implementation |
| 18 | Jaideep Visave | 2025 | AI Transparency in Emergency Management | Trust-building strategies for AI systems | Improved accountability in crisis response |

## 5 Conclusions and Future Studies Recommendation

The present systematic literature review provides evidence that the research area of Explainable AI (XAI) is gaining considerable momentum in safety-sensitive applications, which include healthcare, finance, criminal justice, cybersecurity, industrial controls and environmental systems. The analysed papers unanimously believe that interpretability and transparency are key to trust-building, enhanced accountability and human-AI system collaboration in life-critical systems. Recent XAI approaches such as design of interpretable models, post-hoc techniques, such as LIME and SHAP, counterfactual explanations and Subsequent large language model (LLM) text-based explanations have demonstrated potential to improve user perception, and model decision quality. Nevertheless, there are still permanent problems. They comprise the horizontal scalability of complex high-dimensional explanation approaches, the balance of accuracy versus intelligibility, the lack of standardized assessment frameworks and empirical data about the long-term consequences of XAI on trust and decision-making. Ethical aspects like bias mitigation, protection of privacy or defence of adversarial misuse of explanations, are not systematically covered in those applications. In addition, studies specific to particular sectors demonstrate that trust is not necessarily synonymous with explainability; reliability, validation and fit with user demands may be of equal importance. These results point out the importance of user-centred, context-aware explanations creation, strong benchmarking regimes, and cross-application traits to make XAI systems useful and trustworthy in missions-centric surroundings.

### 5.1 Future Studies Recommendations

1. Create Standard Assessment Models: Design measures of explanation quality that will be accepted universally so as to facilitate balanced

comparisons among techniques and/or across applications.

2. Improve Scalable and Efficient XAI Methods: Design optimizations and hybrid methods that are algorithmically novel and computationally expensive but remain interpretable when scaled to large-scale AI models across domains such as autonomous vehicles, medical imaging, and industrial Internet of Things.

3. Led Longitudinal and Real-World Impact Studies: Study the longer-term impacts of XAI on user trust, correctness of the decision, and the take-up of the system in practice-used conditions, as opposed to the laboratory conditions.

4. Incorporate Human-Centred Design Ideas: Collaboratively develop explanations with end-users and domain professionals to secure agreements in the thinking, decrease cognitive overloads, and overcome biases their automation bias or confirmation bias.

5. Increase Domain related Safety and Reliability Approaches: Extend areas of research beyond trust to cover areas of safety validation, error detection and integrity checks especially in areas critical in safety.

6. Manage Ethical, Privacy and Security Threats: Innovate solutions that can offer transparency, but at the same time do not expose sensitive data, proprietary logic or towards vulnerabilities that might be exploited.

7. Read about Cross-Domain Adaptability: Explore the ways XAI frameworks can be normalized across industries, regulatory landscapes, and cultures but still remain interpretable and successful.

8. Use Large Language Models Responsibility: Research LLM-based explanation systems so that natural language accounts are (practically) loyal to supporting model logic and that risk of hallucinations or misleading explanations are reduced.

*References:*

[1] Markus, A. F., Kors, J. A., & Rijnbeek, P. R. (2020). The role of explainability in creating trustworthy artificial intelligence for health care.

[2] Wysocki, O., Davies, J. K., Vigo, M., Armstrong, A. C., Landers, D., Lee, R., & Freitas, A. (2022). Assessing the communication gap between AI models and healthcare professionals: explainability, utility and trust in AI-driven clinical decision-making.

[3] Fehr, et al. (2025). A Design Framework for Operationalizing Trustworthy Artificial Intelligence in Healthcare: Requirements, Trade-offs and Challenges for its Clinical Adoption.

[4] Saeed, W., & Omlin, C. (2021). Explainable AI (XAI): A Systematic Meta-Survey of Current Challenges and Future Opportunities.

[5] Tiwari, S., Sresth, V., & Srivastava, A. (2023). The Impact of Explainable AI on User Trust in Autonomous Cyber Defense.

[6] Soares, et al. (2020). Prototype-based XAI for autonomous vehicle awareness. *WIREs Data Mining & Knowledge Discovery*.

[7] Gu, et al. (2023). *Air pollution forecasting with deep neural networks and auto-regressive models*. In MDPI Appl.Sci

[8] Chinnaraju, A. (2025). Explainable AI (XAI) for trustworthy and transparent decision-making: A theoretical framework for AI interpretability. *World Journal of Advanced Engineering Technology and Sciences, 14*(03), 170–207. https://doi.org/10.30574/wjaets.2025.14.3.0106.

[9] Adebayo, A. S., Ajayi, O. O., & Chukwurah, N. (2024). Explainable AI in Robotics: A Critical Review and Implementation Strategies for Transparent Decision-Making. *FMR-2025-1-004.1.pdf*, https://doi.org/10.54660/.IJFMR.2024.5.1.26-32

[10] Wiratsin, I.-O., & Ragkhitwetsagul, C. (2024). Effectiveness of Explainable Artificial Intelligence (XAI) Techniques for Improving Human Trust in Machine Learning Models: A Systematic LiteratureReview.https://doi.org/10.1109/ACCESS.2024.0429000.

[11] Kalasampath, K., Spoorthi, K. N., Sajeev, S., Kuppa, S. S., Ajay, K., & Maruthamuthu, A. (2025). A Literature Review on Applications of Explainable Artificial Intelligence (XAI). *IEEE Access*. https://doi.org/10.1109/ACCESS.2025.3546681.

[12] Ayokunle, Micheal Akinsiku. (2025). *Literature Review on Explainable Artificial Intelligence (XAI): Techniques, Tools, and Applications. Tech-Sphere Journal for Pure and Applied Sciences, 2 (1), 1–13. https://doi.org/10.5281/zenodo.15870683*

[13] Jean, G. (2024). Explainable AI for Trust: Developing transparent trust evaluation models that can provide clear explanations for their

assessments. *Explainable+AI+for+TrustDeveloping+transparent+trust+evaluation+models+that+can+provide+clear+explanations+for+their+assessments.pdf*.

[14] Hoenig, A., Roy, K., Acquaah, Y. T., Yi, S., & Desai, S. S. (2024). Explainable AI for Cyber-Physical Systems: Issues and Challenges. *Explainable_AI_for_Cyber Physical_Systems_Issues_and_Challenges.pdf*. https://doi.org/10.1109/ACCESS.2024.3395444.

[15] Ersöz, B., Sağıroğlu, Ş., & Bülbül, H. İ. (2022, September). *A Short Review on Explainable Artificial Intelligence in Renewable Energy and Resources*. In Proceedings of the IEEE 11th International Conference on Renewable Energy Research and Applications (ICRERA)**.** https://doi.org/10.1109/ICRERA55966.2022.9922870

[16] Alexander, D., & Aaron, M. (2025). Explainable AI in cybersecurity: Enhancing transparency and trust. *Proceedings of the International Conference on Cybersecurity and Artificial Intelligence*, 45–53.

[17] Visave, J. (2025). Transparency in AI for emergency management: building trust and accountability. *AI Ethics, 5*, 3967–3980. https://doi.org/10.1007/s43681-025-00692-x

[18] Kanagarla, K. P. B. (2024). Explainable AI in data analytics: Enhancing transparency and trust in complex machine learning models. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.5012468

[19] Reddy, A. K., Bojja, S. G. R., Thota, S., Chitta, S., & Saini, V. (2024). Bridging AI and Human Understanding: Interpretable Deep Learning in Practice. *Journal of Informatics Education and Research, 4*(3), 3705–3720. http://jier.org

[20] Duarte, R. de B., Correia, F., Arriaga, P., & Paiva, A. (2023). AI Trust: Can Explainable AI Enhance Warranted Trust? *Human Behavior and Emerging Technologies, 2023*, 1–12. https://doi.org/10.1155/2023/4637678

[21] Jiang, H., & Senge, E. (2021). *On Two XAI Cultures: A Case Study of Non-technical Explanations in Deployed AI System*.

[22] Haque, A. K. M. B., Islam, A. K. M. N., & Mikalef, P. (2023). NOTION OF EXPLAINABLE ARTIFICIAL INTELLIGENCE – AN EMPIRICAL INVESTIGATION FROM A USER'S PERSPECTIVE. *ECIS 2023 Research Papers*, Paper 404.

[23] Bach, T. A., Khan, A., Hallock, H., Beltrão, G., & Sousa, S. (2022). A Systematic Literature Review of User Trust in AI-Enabled Systems: An HCI Perspective. *International Journal of Human–Computer Interaction*, 40(5), 1251–1266. https://doi.org/10.1080/10447318.2022.2138826.

[24] Zielinska, K. (2024, May 30). *Designing AI Clinical Decision Support Systems with a Human-Centric Usability Focus: Designs AI-driven clinical decision support systems with a focus on user-centered design principles to enhance usability and adoption*. Journal of AI-Assisted Scientific Discovery**,** 4 (1).

[25] Chandra et al. (2022). *A systematic review of intelligent support systems for strategic decision-making using human-AI interaction in enterprise platforms*.

[26] Muley, A., Muzumdar, P., Kurian, G., & Basyal, G. P. (2023). *Risk of AI in Healthcare: A Comprehensive Literature Review and Study Framework*.

[27] Ferrara, E. (2023). *Fairness And Bias in Artificial Intelligence: A Brief Survey of Sources, Impacts, And Mitigation Strategies*.

[28] Reuters (2024). *Legal transparency in AI finance: facing the accountability dilemma in digital decision-making*.

[29] Talukder, K. A., & Shompa, T. F. (2024). *Artificial Intelligence in Criminal Justice Management: A Systematic Literature Review*. Journal of Machine Learning.

[30] Riega-Virú, Y. B., Soto, N. E., Salas, J. M., Natividad, P., Salas-Riega, J. L., & Nilupú-Moreno,K. (2024). *Artificial Intelligence and Criminal Justice: A systematic review*. LACCEI. proceedings.laccei.org

[31] Noiret, S., Lumetzberger, J., & Kampel, M. (2022). *Bias and Fairness in Computer Vision Applications of the Criminal Justice System*.

[32] Kennedy, R., Tiede, L., Austin, A., & Ismael, K. (2025). *Law Enforcement and Legal Professionals' Trust in Algorithms*. SAGE Journals

[33] Chen, Z. Z., Ma, J., Zhang, X., Hao, N., Yan, A., Nourbakhsh, A., et al. (2024). *A Survey on Large Language Models for Critical Societal Domains: Finance, Healthcare, and Law*.

[34] Li, X., Xiong, H., Liu, J., et al. (2021). *Interpretable Deep Learning: Interpretation, Interpretability, Trustworthiness, and Beyond*.

[35] Belle, V., & Papantonis, I. (2020). *Principles and Practice of Explainable Machine Learning*.

[36] Sokol, K., & Flach, P. (2020). *Explainability Fact Sheets: A Framework for Systematic Assessment of Explainable Approaches*. In *Proceedings of the ACM FAT* '20*
[37] Dieber, J., & Kirrane, S. (2022). *A novel model usability evaluation framework (MUsE) for explainable artificial intelligence*. *Information Fusion.*
[38] Gao et al. (2024) Open-source transparency concerns (2024).
[39] Tufail, M. (2024). *AI-Driven Modern Cybersecurity Approach: A Systematic Literature Review*.
[40] Mendes, C., & Rios, T. N. (2023). *Explainable Artificial Intelligence and Cybersecurity.*
[41] Kimanzi, R., Kimanga, P., Cherori, D., & Gikunda, P. K. (2024). *Deep Learning Algorithms Used in Intrusion Detection Systems.*
[42] Jacob, S. L., & Habibullah, P. S. (2024). *A Systematic Analysis and Review on Intrusion Detection Systems Using Machine Learning and Deep Learning Algorithms*.
[43] Richards, E. (2024). *Deep Learning Techniques for Intrusion Detection Systems: A Comparative Study of Accuracy and Efficiency*.
[44] ScienceDirect (2022). *An intelligent and efficient network intrusion detection system using deep learning*. Computers and Electrical Engineering.
[45] Agrawal, S., et al. (2021). *Federated Learning for Intrusion Detection System: Concepts,Challenges and Future Directions*.
[46] Mumuni, F., & Mumuni, A. (2025). Explainable Artificial Intelligence (XAI): From Inherent Explainability to Large Language Models. *XAI FROM INHERENT EXPLAINABILITY TO LARGE LANGUAGE MODELS.pdf*.
[47] Zytek, A., Pido, S., Alnegheimish, S., Berti-Equille, L., & Veeramachaneni, K. (2024). *Explingo: Explaining AI Predictions using Large Language Models.*
[48] Paul, A. L. (2025). *Improving Explainability in Large Language Models (LLMs)*. ResearchGate.
[49] Preprints.org (2025). *Exploring Explainability in Large Language Models*. Preprints.
[50] Mohyuddin, G., Khan, M. A., Haseeb, A., Mahpara, S., Waseem, M., & Saleh, A. M. (2024). Evaluation of machine learning approaches for precision farming in smart agriculture system: A comprehensive review. *IEEE Access, 12*, 60155–60184. https://doi.org/10.1109/ACCESS.2024.3390581

**Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)**

The authors equally contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

**Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself**

No funding was received for conducting this study.

**Conflict of Interest**

The authors have no conflicts of interest to declare that are relevant to the content of this article.