# XAI-Compliance-by-Design: A Modular Framework for GDPR- and AI Act-Aligned Decision Transparency in High-Risk AI Systems

Antonio Goncalves * and Anacleto Correia

Posted Date: 1 December 2025

doi: 10.20944/preprints202512.0062.v1

Keywords: cybersecurity; privacy; Explainable Artificial Intelligence (XAI); GDPR compliance; AI Act; trustworthy MLOps; AI governance; algorithmic accountability; high-risk AI systems

 doi:10.20944/preprints202512.0062.v1



*Article*

# XAI-Compliance-by-Design: A Modular Framework for GDPR- and AI Act-Aligned Decision Transparency in High-Risk AI Systems

**Antonio Goncalves** [1,*] **and Anacleto Correia** [2]

CINAV; 2810-001 Almada
* Correspondence: agoncalvesLX@gmail.com

**Abstract**

High-risk Artificial Intelligence (AI) systems deployed in cybersecurity and privacy-critical contexts must satisfy not only demanding performance targets but also stringent obligations for transparency, accountability and human oversight under the General Data Protection Regulation (GDPR) and the Artificial Intelligence Act (AI Act). Existing approaches often treat these concerns in isolation: explainable AI (XAI) methods are added ad hoc to machine learning pipelines, while governance and regulatory frameworks remain largely conceptual and weakly connected to the concrete artefacts produced in practice. This article proposes *XAI-Compliance-by-Design*, a modular framework that integrates XAI techniques, compliance-by-design principles and trustworthy Machine Learning Operations (MLOps) practices into a unified architecture for high-risk AI systems in cybersecurity and privacy domains. The framework follows a dual-flow design that couples an upstream technical pipeline (data, model, explanation and monitoring) with a downstream governance pipeline (policy, oversight, audit and decision-making), orchestrated by a Compliance-by-Design Engine and a technical–regulatory correspondence matrix aligned with the GDPR, the AI Act and ISO/IEC 42001. The framework is instantiated and evaluated through an end-to-end, Python-based proof of concept using a synthetic, intrusion detection system (IDS)-inspired anomaly detection scenario with a Random Forest classifier, SHAP and LIME explanations, drift indicators and tamper-evident evidence bundles and decision dossiers. The results show that, even in a modest, toy setting with limited predictive performance, the framework systematically produces verifiable artefacts that support auditability and accountability across the model lifecycle. By linking explanation reports, drift statistics and compliance logs to concrete regulatory provisions, the approach illustrates how organisations operating high-risk AI for cybersecurity and privacy can move from model-centric optimisation to evidence-centric governance. The article discusses how the proposed framework can be generalised to real-world high-risk AI applications, contributing to the operationalisation of European digital sovereignty in AI governance.

**Keywords:** cybersecurity; privacy; Explainable Artificial Intelligence (XAI); GDPR compliance; AI Act; trustworthy MLOps; AI governance; algorithmic accountability; high-risk AI systems

## 1. Introduction

### *1.1. Motivation and Context*

The increasing deployment of Artificial Intelligence (AI) in high-risk decision-making contexts—including healthcare, finance, critical infrastructures and public administration—has intensified demands for transparency, accountability and effective human oversight throughout the algorithmic lifecycle.

Within the European regulatory landscape, the combined requirements of the General Data Protection Regulation (GDPR) [1] and the Artificial Intelligence Act (AI Act) [2] impose stringent obligations on how automated decision systems must be designed, documented and audited, requiring

 doi:10.20944/preprints202512.0062.v1



organisations to treat transparency, accountability and human oversight as first-class design constraints rather than post hoc add-ons.

From a technical perspective, the rapid evolution of machine learning has exposed a persistent gap between mainstream engineering practices and legal–regulatory expectations.

The literature on Explainable Artificial Intelligence (XAI) has consistently emphasised interpretability, traceability and auditability as central pillars of algorithmic trustworthiness [3,4], yet these principles are rarely translated into concrete artefacts and metrics that can be inspected by data protection officers, auditors or regulators. In many production environments, explanation mechanisms remain ad hoc, unversioned and weakly connected to formal compliance processes, hindering the systematic demonstration of conformity with GDPR and AI Act requirements [5].

Security- and privacy-critical domains form a particularly demanding subset of high-risk AI applications. Examples include fraud detection, anomaly detection in critical infrastructures, cyberthreat detection in network traffic and risk scoring in electronic public services. In such settings, engineering teams must ensure not only accurate and robust models but also operational evidence that these models behave in a stable, explainable and accountable manner over time. This combination of operational criticality and regulatory scrutiny calls for integrated frameworks that embed XAI, compliance-by-design and trustworthy Machine Learning Operations (MLOps) principles directly into the lifecycle of high-risk AI systems.

Existing approaches typically address these concerns in a fragmented manner. XAI methods are often implemented as add-ons to existing pipelines; governance frameworks remain largely conceptual; and regulatory analyses frequently lack explicit links to the artefacts actually produced by machine learning workflows. There remains a lack of a unified, operational framework that connects technical metrics and artefacts—such as models, explanations, logs and drift indicators—to concrete regulatory requirements in a way that is both implementable with mainstream tools and auditable in practice.

This need is especially evident in cybersecurity operations, such as security operations centres (SOCs) and intrusion detection systems (IDS), where teams must justify alerts and interventions in environments subject to strict privacy and logging constraints.

### *1.2. Problem Statement*

This work addresses the following central question: *how can we design and implement an operational framework that integrates Explainability, Compliance-by-Design and trustworthy MLOps into high-risk AI systems in a way that produces verifiable evidence of conformity with the GDPR and the AI Act?*

Current solutions often focus primarily on model performance and ad hoc explanations, or on high-level governance principles that are not operationally implemented in machine learning pipelines. There is a lack of end-to-end architectures that (i) explicitly map technical artefacts to regulatory requirements, (ii) provide systematic compliance logging and tamper-evident evidence bundles, and (iii) can be instantiated in realistic, security- and privacy-critical scenarios without requiring bespoke tooling.

Addressing this gap is essential to operationalising European regulatory obligations in a technically grounded manner and supporting European digital sovereignty in the governance of high-risk AI systems.

### *1.3. Research Objectives and Contributions*

The overarching goal of this article is to propose and validate a modular *XAI-Compliance-by-Design* framework that bridges the gap between regulatory requirements and machine learning practice in high-risk AI systems. This research is guided by the following questions (RQs):

- **RQ1:** How can XAI techniques, compliance-by-design principles and trustworthy MLOps practices be integrated into a single modular framework for high-risk AI systems?

 doi:10.20944/preprints202512.0062.v1



- **RQ2:** To what extent can such a framework produce concrete, verifiable artefacts—such as models, explanations, logs and evidence bundles—that support GDPR- and AI Act-aligned auditability and accountability?
- **RQ3:** How does the proposed framework behave when instantiated in an anomaly detection scenario using a synthetic, security-relevant tabular dataset, in terms of predictive performance, explainability and drift monitoring?

In response to these questions, the article contributes:

1. A conceptual *XAI-Compliance-by-Design* framework linking the lifecycle of high-risk AI systems to regulatory requirements from the GDPR, the AI Act and ISO/IEC 42001, emphasising traceability, oversight and risk management.
2. A modular, MLOps-oriented architecture integrating data preprocessing, model training, explainability, drift monitoring and compliance logging, designed to be implementable with widely used open-source tools.
3. A technical–regulatory correspondence matrix mapping specific metrics and artefacts—such as SHAP reports, drift statistics, model lineage logs and evidence bundles—to relevant legal and standardisation provisions.
4. An end-to-end proof-of-concept implementation in Python, instantiated on a synthetic, IDS-inspired anomaly detection dataset using a Random Forest classifier combined with SHAP and LIME explanations, producing versioned models, explanation artefacts, drift indicators and tamper-evident evidence bundles. This implementation serves as an illustrative toy example to demonstrate implementability rather than to optimise intrusion detection performance.
5. An empirical assessment of model performance, global and local explainability and stability under dataset shift and distributional drift, discussing the implications for trustworthy MLOps, regulatory governance and European digital sovereignty.

Together, these contributions demonstrate that XAI, regulatory alignment and compliance documentation can be embedded directly into the technical fabric of high-risk AI pipelines rather than treated as external or purely conceptual layers.

### *1.4. Structure of the Paper*

The remainder of this paper is structured as follows. Section 2 reviews the state of the art in XAI, compliance-by-design and AI accountability, identifying the conceptual and operational gaps that motivate the proposed framework. Section 3 describes the research design, the framework-oriented methodology and the illustrative case study. Section 4 presents the *XAI-Compliance-by-Design* framework, detailing its architectural logic, functional layers and technical–regulatory correspondence matrix. Section 5 reports the implementation details and experimental validation, including model performance, feature importance analysis, global explainability and drift monitoring. Section 6 discusses the technical, regulatory and strategic implications of the results, and Section 7 concludes the paper and outlines directions for future research.

## 2. Related Work and State of the Art

Recent research in Explainable Artificial Intelligence (XAI) has increasingly focused on the pressing needs of cybersecurity and privacy-critical applications. These demands are particularly prominent within the context of the rapidly evolving European regulatory environment.

This section provides a critical analysis of the technological, scientific, and legal progress that defines the pillars of explainability, auditability, and compliance-by-design required for trustworthy, high-risk AI systems.

The discussion is structured into five subsections: (1) an overview of the foundations of XAI; (2) an analysis of the interplay between regulatory and technical dimensions; (3) a comparative review of existing frameworks; (4) an examination of legal and ethical foundations; and (5) the conceptual positioning of the current contribution.

 doi:10.20944/preprints202512.0062.v1



Explainable Artificial Intelligence (XAI) has emerged in response to the growing demand for transparency in algorithmic outputs generated by complex, opaque machine learning models [6]. Foundational studies, such as Doshi-Velez and Kim [4], have defined interpretability as a measurable property within the broader algorithmic lifecycle. Interpretability may involve intrinsic transparency, where simple models are inherently explainable, or post-hoc techniques designed to clarify decisions made by black-box systems [3,6].

Techniques including LIME, SHAP, TreeSHAP, and surrogate models provide both local and global explanations to support human-understandable justifications for automated decisions. However, these approaches may be affected by instability, bias, and high computational cost [3,4]. In practical settings, explainability has become fundamental for fostering institutional trust, especially in cybersecurity and privacy-critical applications. Studies highlight that explanation quality hinges on robustness, stability, and contextual relevance—all attributes vital for trustworthy, auditable MLOps pipelines [7].

While these contributions lay the scientific groundwork for linking explainable AI with regulatory obligations, most foundational studies do not yet provide standardized, audit-ready compliance metrics. Addressing this challenge is key to enabling compliance-by-design in high-risk AI systems consistent with evolving regulatory expectations.

### *2.1. Compliance-by-Design and AI Accountability*

Compliance-by-design strengthens the integration of regulatory requirements into the entire algorithmic lifecycle, embedding legal and ethical controls within the engineering of AI systems rather than relying solely on retrospective verification [8]. This paradigm, rooted in privacy-by-design under Article 25 of GDPR [1], expands to include transparency, human oversight, risk management, and auditability in a unified operational framework.

Current literature provides limited guidance on systematically implementing these controls in large-scale, continuously evolving AI pipelines. Best practices increasingly require translating process controls into technical artefacts—such as automated risk assessments, decision logs, and traceable audit trails—that are continuously updated and legally mapped to GDPR and AI Act obligations.

As a result, compliance-by-design initiatives often remain theoretical, lacking proven workflows for evidence curation, versioning, and traceability. The ISO/IEC 42001 standard strengthens the discipline of audit management by mandating explicit controls, ongoing documentation, and the generation of audit-ready regulatory evidence—including risk categorization, bias testing, and incident reporting throughout the full model lifecycle.

Continuous accountability requires automated management, documentation, and monitoring of AI behaviours, with clear roles allocated and readiness for audit. These elements are critical for trustworthy AI governance, particularly in high-risk cybersecurity and privacy applications.

### *2.2. Comparative Analysis of Existing Frameworks*

Several frameworks have sought to align explainability with regulatory requirements. However, most exhibit operational limitations, including the absence of standardized processes for generating, versioning, and maintaining compliance evidence—such as decision logs, model documentation, and audit trails—across the entire AI lifecycle.

Table 1 summarises the most relevant contributions from 2020–2025.

 doi:10.20944/preprints202512.0062.v1



**Table 1.** Comparison of regulatory-relevant XAI frameworks.

| Framework | Key characteristics | Regulatory alignment |
|---|---|---|
| Arrieta et al. (2020) [3] | Taxonomy of XAI techniques; fidelity and stability analysis. | Supports transparency; lacks explicit regulatory mapping. |
| Chhetri et al. (2022) [8] | Semantic modelling for automated GDPR conformance. | Implements Data Protection by Design. |
| Liao et al. (2022) [9] | Algorithmic auditing and explainability controls. | Covers decision provenance and human oversight. |
| Kabir et al. (2025) [10] | Organisational trust and governance perspective. | Identifies measurable accountability metrics. |
| Kostopoulos et al. (2024) [11] | Operational transparency in decision-support systems. | Partially aligned with AI Act transparency obligations. |
| Islam et al. (2024) [7] | Unified evaluation framework for explanations. | Offers audit-relevant metrics; lacks full legal integration. |
| Longo et al. (2024) [12] | XAI 2.0 research manifesto; interdisciplinary challenges. | Addresses comprehensibility and transparency. |
| Pinto (2024) [13] | Standardised empirical evaluation of explanations. | Facilitates audit standardisation. |
| Pavlidis (2025) [14] | Application of XAI within the AI Act framework. | Explicitly AI Act–oriented. |

Operationally robust frameworks must address crucial compliance pillars: automated lineage tracking, versioning of code and models, continuous drift monitoring, and granular, audit-grade trails that meet legal standards for high-risk AI under the EU AI Act [2].

Despite these conceptual contributions, only a minority of frameworks currently offer machine-verifiable mechanisms for continuous, MLOps-native compliance—such as automated lineage tracking, versioned decision logs, model cards, and tamper-evident audit trails. Recent standards, including EU AI Act Article 96 and ISO/IEC 42001 [15], now require compliance artifacts to be continuously retrievable and mapped to explicit legal controls. However, most reviewed frameworks do not yet reach these operational and legal benchmarks.

The solution presented in this article directly addresses these deficiencies by integrating real-time evidence generation, robust versioning, and comprehensive model lifecycle governance into a unified, auditable approach to AI compliance.

### *2.3. Regulatory and Ethical Foundations: GDPR, AI Act, and Digital Sovereignty*

The European regulatory framework provides the foundation for trustworthy and accountable AI operations. The GDPR [1] establishes principles of transparency, accountability and lawfulness applicable to automated decision-making systems processing personal data. The AI Act [2] reinforces this structure through a risk-based classification of AI systems and mandatory requirements for documentation, testing, traceability and human oversight in high-risk applications.

ISO/IEC 42001:2023 [15] extends this governance architecture by defining a structured management system for AI, detailing controls for oversight, risk management, documentation and continuous monitoring aligned with both the GDPR and the AI Act.

Recent work by Ahangar et al. [5] and Lozano-Murcia et al. [16] emphasises that European digital sovereignty depends on verifiable technical infrastructures capable of producing trustworthy, auditable and reproducible evidence—precisely the type of evidence targeted by the framework proposed in this article.

### *2.4. Research Gap and Conceptual Positioning*

A persistent gap across the literature concerns the limited operationalisation of explainability metrics—such as fidelity, stability and comprehensibility—as measurable indicators of regulatory alignment [3]. Existing frameworks tend to focus either on technical explainability or on legal obligations but rarely offer integrated, audit-ready mechanisms that link the two domains [8].

 doi:10.20944/preprints202512.0062.v1



This disconnection poses practical challenges for engineering and compliance teams, who must translate technical artefacts into regulatory evidence that can be examined by auditors, data protection officers and supervisory authorities.

The present work addresses this gap by proposing a modular *XAI-Compliance-by-Design* framework that connects explainability metrics to concrete regulatory requirements and produces verifiable evidence bundles integrated directly into the MLOps pipeline. This enables continuous, reproducible and audit-ready compliance throughout the lifecycle of high-risk AI systems, aligning technical, organisational and regulatory dimensions in support of European digital sovereignty.

## 3. Methodology

This section presents the methodological approach adopted to design, formalise and operationalise the proposed *XAI-Compliance-by-Design* framework. The primary focus is the construction of a general, reusable framework that can be instantiated across multiple high-risk AI contexts, rather than the optimisation of any particular machine learning model or domain-specific use case. The anomaly detection scenario described later in this section is used as an illustrative synthetic example, whose sole purpose is to demonstrate the implementability of the framework and the generation of audit-ready artefacts.

Accordingly, the main evaluation criterion in this work is not improved intrusion detection performance, but the ability of the proposed pipeline to operationalise regulatory obligations under the GDPR, the AI Act and ISO/IEC 42001 in a traceable and auditable manner. The synthetic IDS-like scenario is deliberately kept simple to isolate the contribution of the framework and its evidence-generation flow, avoiding domain-specific optimisations that could obscure the compliance-by-design mechanisms that are central to this study.

The methodology combines three mutually reinforcing components: (i) a regulatory-informed conceptual model grounded in the GDPR, the AI Act and ISO/IEC 42001; (ii) a Design Science Research (DSR) process focused on the construction and evaluation of verifiable artefacts; and (iii) an operational MLOps pipeline that implements the framework and produces technical and regulatory evidence. The complete implementation (Python code, configuration files and executable notebooks) is provided as supplementary material.

Sugestão de melhoria, já pronta para substituir no LaTeX, ajustada para o foco em cibersegurança/privacidade e AI compliance (estilo JCP/MDPI):

### *3.1. Framework-Oriented Methodology and MLOps Pipeline*

The framework is operationalised through a generic, compliance-oriented MLOps pipeline designed to be applicable across different high-risk AI domains, with a particular focus on cybersecurity and privacy-sensitive settings. The pipeline is organised into stages that mirror the lifecycle of an AI system, from data handling to deployment-oriented evidence generation, and each stage is instrumented with compliance logging so that technical events can be systematically traced back to regulatory requirements.

At a high level, the pipeline comprises:

- **Environment configuration and context registration:** initialisation of the execution environment, creation of working directories, configuration of random seeds and registration of key parameters (e.g., data sources, model family, hyperparameter ranges) in a compliance log. Each execution is assigned a unique identifier (`RUN_ID`) that links all subsequent artefacts, including software versions and configuration files.
- **Data handling and preprocessing:** ingestion or generation of data, separation into features and target variables, definition of numerical and categorical attributes and configuration of preprocessing steps (e.g., *ColumnTransformer*, scaling, encoding). The resulting schema and data statistics are recorded to support documentation obligations under the AI Act and ISO/IEC 42001 regarding data quality, representativeness and preprocessing.

 doi:10.20944/preprints202512.0062.v1



- **Model training and validation:** training of a classification model encapsulated in a `scikit-learn`-based pipeline, with standard train–test splitting and computation of performance metrics (accuracy, precision, recall, F1-score and AUC-ROC). The choice of model family (here, a Random Forest) is illustrative and not essential to the framework, which can be instantiated with alternative classifiers that fit the same pipeline structure.
- **Explainability and drift monitoring:** generation of global and local explanations (SHAP, LIME) and computation of basic drift indicators on held-out or temporally segmented data, in order to demonstrate how explainability and monitoring artefacts are integrated into the compliance workflow. These artefacts are later used as inputs for audit-oriented documentation, model risk analysis and human oversight.
- **Evidence bundle construction:** aggregation of models, metrics, explanation outputs, drift statistics and compliance logs into structured evidence bundles (e.g., JSON manifests and directory structures) that can be inspected by auditors or regulators. These bundles are designed to be directly reusable as building blocks for technical documentation dossiers and conformity assessment under the AI Act.

The pipeline thus serves as a vehicle for instantiating the framework; its structure and logging mechanisms are designed to be transferable to other application domains beyond the illustrative anomaly detection scenario, provided that domain-specific data and models can be mapped to the same evidence and compliance-logging pattern.

### *3.2. Illustrative Case Study: Synthetic IDS-like Scenario*

To demonstrate the practical instantiation of the framework in a cybersecurity-relevant setting, a synthetic network anomaly detection scenario is used as an illustrative example. This case study has a purely demonstrative function: it shows how the framework can be implemented end-to-end and how the corresponding artefacts are generated, versioned and logged, without making domain-level claims about intrusion detection performance.

In this scenario, synthetic network traffic is generated to emulate typical intrusion detection system (IDS) datasets. The generated dataset comprises:

- **Numerical and categorical features:** continuous predictors (e.g., connection duration, packet and byte counts) and categorical variables (e.g., protocol type, service, flags), processed via a *ColumnTransformer* with *OneHotEncoder* for categorical attributes and passthrough for numerical attributes.
- **Binary target variable:** a label *labels* representing *normal* and *attack* traffic, used solely to demonstrate how the framework manages supervised classification tasks in an IDS-like context.
- **Class imbalance:** a minority proportion of *attack* events (around 20%), echoing the typical imbalance found in many operational network settings, without claiming to reproduce any specific real-world environment.

The dataset is frozen and stored in a persistent data layer together with metadata describing its dimensionality, feature types and class distribution. These elements are registered in the compliance log to provide an auditable record of the data configuration used in this particular instantiation and to support reconstruction of the experimental setup.

The simplicity of this synthetic scenario is intentional: it removes confounding factors associated with complex, real-world SOC environments and allows the evaluation to focus on whether the framework and pipeline produce the expected lineage records, explanation artefacts and compliance logs. The empirical setting should therefore be interpreted as a minimal, controlled environment for exercising the compliance-by-design machinery, rather than as an attempt to advance the state of the art in intrusion detection.

 doi:10.20944/preprints202512.0062.v1



### 3.3. Explainability Layer: SHAP and LIME

The explainability layer is designed to be model- and domain-agnostic, and the choice of SHAP and LIME reflects their maturity and extensive use in XAI practice for tabular, security- and privacy-relevant data. In the illustrative scenario, they are applied as follows:

- **SHAP:** Shapley values are computed via a *TreeExplainer* on a representative subset of the transformed dataset. The outputs include global summary plots and feature importance statistics, which demonstrate how the framework can generate explanation artefacts suitable for incorporation into audit-ready evidence bundles and for assessing properties such as fidelity and stability.
- **LIME:** Instance-level explanations are generated using a *LimeTabularExplainer*, focusing on selected cases (e.g., false positives and false negatives). The goal is to show how local explanations can be captured, stored and linked to compliance events, supporting human oversight and documentation of decision rationales rather than analysing any specific operational scenario.

All explanation-related operations—including configuration parameters, sampling strategies and file locations—are logged as part of the compliance record, enabling reproducibility, systematic re-analysis and audit re-execution across different instantiations of the framework.

### 3.4. Evaluation Criteria and Assessment Model

The evaluation focuses on assessing the framework and its artefacts rather than optimising the underlying illustrative model. Three complementary dimensions are considered:

1. **Model performance (illustrative):** standard metrics such as accuracy, precision, recall, F1-score and AUC-ROC are computed to confirm that the toy model behaves in a plausible manner. These metrics are reported to contextualise the explanation and compliance artefacts and to provide a basic characterisation of predictive behaviour, not as the primary contribution of the work.
2. **Explanation properties:** fidelity, stability and comprehensibility are assessed qualitatively and, where applicable, via metrics inspired by recent unified evaluation frameworks for XAI [7,13]. The objective is to verify that the framework is capable of producing explanations that can be systematically documented, revisited and compared across runs, rather than to exhaustively benchmark XAI techniques.
3. **Compliance and governance indicators:** coverage of regulatory obligations (percentage of GDPR, AI Act and ISO/IEC 42001 requirements mapped to technical controls in the technical–regulatory correspondence matrix), completeness of evidence bundles and the ability to reconstruct, from the compliance log, the full data–model–explanation–decision lineage for a given `RUN_ID`. These indicators capture the extent to which the pipeline supports continuous, audit-ready governance of high-risk AI systems.

These criteria are used to determine whether the framework fulfils its design objectives of linking technical artefacts to regulatory requirements in an operational, audit-ready manner. In other words, the success of the study is primarily measured by the extent to which the pipeline can generate coherent and reusable evidence of conformity with GDPR, AI Act and ISO/IEC 42001 obligations, with model performance playing a secondary, contextual role.

### 3.5. Limitations, Ethical Considerations, and Reproducibility

The methodological choices entail several limitations. The use of synthetic data deliberately constrains the external validity of the empirical results: the anomaly detection scenario is a simplified example and does not aim to replicate the full complexity, variability or adversarial characteristics of real-world network environments. Similarly, the selection of a Random Forest classifier is illustrative and does not preclude applying the framework to other model families or domains, including deep learning and hybrid architectures in cybersecurity and privacy-sensitive contexts.

 doi:10.20944/preprints202512.0062.v1



From an ethical standpoint, the framework recognises that explainability can be misused to provide a veneer of transparency for systems that remain fundamentally opaque or that operate in settings with significant power asymmetries. To mitigate this risk, the methodology emphasises: (i) explicit documentation of modelling assumptions and trade-offs; (ii) systematic recording of all relevant artefacts and decisions; and (iii) alignment with the transparency, necessity and proportionality principles embedded in the GDPR and the AI Act, including meaningful human oversight and the possibility to contest automated decisions.

Reproducibility is supported through complete documentation of the pipeline in Python, versioned dependencies, fixed random seeds and structured storage of data, models, explanations, metrics and compliance logs. Execution identifiers (`RUN_ID`) and file paths are consistently registered, enabling other researchers, auditors or regulators to reinstantiate the framework and repeat the illustrative experiments under comparable conditions, or to adapt the same methodology to different high-risk AI domains, subject to their own data protection impact assessments and sector-specific risk analyses.

## 4. Proposed Framework: XAI-Compliance-by-Design

This section presents the modular *XAI-Compliance-by-Design* framework, designed to natively integrate explainability, accountability and continuous regulatory compliance into the lifecycle of high-risk AI systems. The framework explicitly addresses the disconnect identified in Section 2 between technical explanation metrics—such as fidelity, stability and comprehensibility—and the legal requirements imposed by the GDPR [1], the AI Act [2] and ISO/IEC 42001:2023 [15].

The framework is structured around two tightly coupled and synchronised flows: an *upstream* technical flow responsible for managing data, models, explanations and technical evidence, and a *downstream* governance flow that translates legal and policy requirements into enforceable technical controls and continuous monitoring [3,8]. At the centre of the architecture, a *Compliance-by-Design Engine* (CDE) orchestrates the alignment between these flows, ensuring that technical metrics and artefacts are systematically mapped to regulatory obligations and that audit-ready evidence is produced throughout the system lifecycle.

### *4.1. Conceptual Overview and Architectural Logic*

The conceptual architecture of the *XAI-Compliance-by-Design* framework is depicted in Figure 1. It is grounded in two complementary flows, mediated by the CDE, that jointly ensure technical robustness and regulatory alignment across the lifecycle of high-risk AI systems.

The **upstream technical flow** represents the operational lifecycle of the AI system: it starts with data acquisition and preparation, proceeds through model training, evaluation and explainability, and culminates in the generation and storage of structured technical evidence (e.g., explanation reports, drift metrics, lineage logs). This flow is *upstream* in the sense that it builds increasingly abstract artefacts from raw data towards higher-level representations of model behaviour and risk.

The **downstream governance flow** models the translation of legal, regulatory and organisational requirements into operational controls. It starts from principles and obligations derived from the GDPR, the AI Act and ISO/IEC 42001, which are then instantiated as concrete policies, monitoring rules and audit procedures. This flow is *downstream* because it propagates requirements from the normative layer towards the technical infrastructure, constraining and guiding system behaviour through policy-as-code, documentation requirements and oversight mechanisms.

The CDE operates as a dynamic orchestrator that synchronises these two flows in real time. It maintains a formal mapping between technical metrics and regulatory indicators (e.g., linking explanation fidelity to transparency obligations, or drift thresholds to post-market monitoring duties) and continuously evaluates whether the artefacts generated by the upstream flow satisfy the expectations encoded in the downstream flow. This includes tracking model lineage, decision provenance and coverage of compliance requirements across builds and deployments.

 doi:10.20944/preprints202512.0062.v1



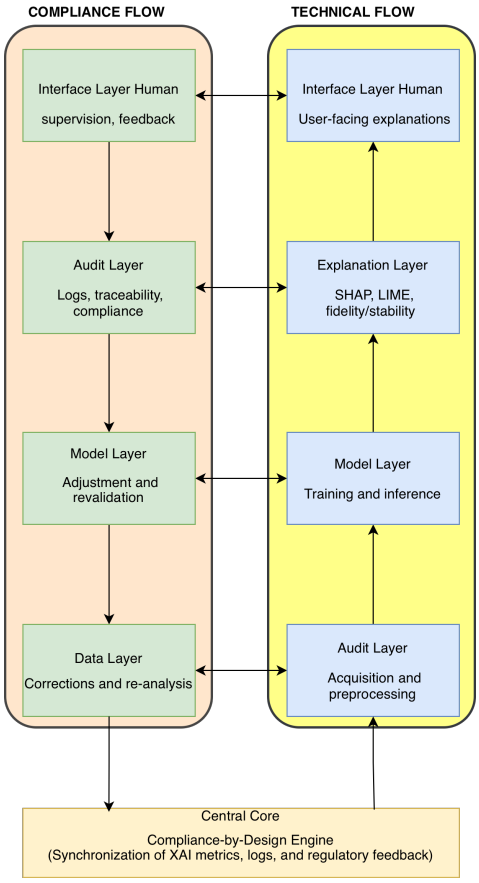


**Figure 1.** Conceptual architecture of the XAI-Compliance-by-Design model.

A key feature of the architecture is the presence of *bidirectional horizontal communication* between homologous modules in the technical and governance flows. For example, audit modules consume explanation reports, drift metrics and logs produced by the technical pipeline, while also issuing alerts, recommendations or updated policies that feed back into model retraining, threshold adjustment or user interface adaptations. This cyclical interaction operationalises bidirectional oversight models advocated in recent AI governance literature and supports continuous alignment between technical practice and evolving regulatory expectations.

Overall, the architecture emphasises modularity, separation of concerns and responsiveness: technical components can evolve (e.g., model families, explainers) without breaking the governance layer, while regulatory changes (e.g., new guidance or standards) can be reflected in the CDE and propagated downstream without requiring a complete redesign of the technical stack.

### *4.2. Functional Layers*

The framework is organised into functional layers that span both the technical and governance flows, ensuring native compatibility between explainability, operational management and regulatory oversight [17–21]. These layers are aligned with the dual pipeline illustrated in Figure 1: the upstream flow emphasises data processing, modelling and explanation, whereas the downstream flow focuses on verification, continuous auditing and human oversight [22,23].

The main functional layers and their relationships with regulatory requirements and technical metrics are as follows:

- **Data Layer:** responsible for data acquisition, preprocessing and data quality auditing, ensuring integrity, provenance and traceability. This layer underpins compliance with GDPR Articles 5 and 25 and supports robust governance practices through data documentation, quality indicators and impact assessment inputs [24,25].

 doi:10.20944/preprints202512.0062.v1



- **Model Layer:** covers model development, validation, deployment and monitoring, with emphasis on traceability, versioning and documentation for accountability and transparency in line with the AI Act and GDPR [1,2]. This layer supports formal revalidation processes whenever technical audits detect performance degradation, emerging risks or material changes in data or use context.
- **Explanation Layer:** implements explanation techniques and metrics—such as SHAP, LIME, fidelity and stability—to provide human-understandable justifications for automated decisions and to generate technical evidence for legal justification duties [26–30]. It directly supports transparency provisions under the GDPR and AI Act and feeds explanation artefacts into the audit and governance processes.
- **Audit Layer:** coordinates continuous evaluation and monitoring of compliance, automating reporting and log management to meet obligations such as responding to access or contestation requests (e.g., GDPR Article 22) and post-market monitoring under the AI Act [22,31,32]. This layer maintains integrity checks and preserves audit trails across the lifecycle, linking technical events to governance decisions.
- **Interface Layer:** provides interactive mechanisms for visualising explanations and supporting human oversight, thereby operationalising meaningful human intervention and accountability [17, 24]. It enables human-in-the-loop review, feedback and override capabilities, which are critical in high-risk AI settings and for demonstrating effective human oversight in conformity assessments.

At the core, the **Compliance-by-Design Engine (CDE)** aggregates XAI metrics, logs and governance feedback, updating compliance indicators and triggering actions (e.g., retraining, policy updates, enhanced monitoring) when thresholds are breached [8,14,31–33]. This central orchestration ensures that compliance is not a one-off activity but a continuous process embedded into the operational fabric of the AI system.

To ground these design choices, Table 2 summarises the architectural principles incorporated into the framework, linking them to seminal references and to their concrete manifestations in the proposed design.

**Table 2.** Architectural principles incorporated into the XAI-Compliance-by-Design framework.

| Principle | Reference | Description | Application in the framework |
|---|---|---|---|
| Modularity | [34,35] | Independent, evolvable components. | Separated functional layers and replaceable modules across the technical and governance flows. |
| Separation of concerns | [36,37] | Each module has a single, well-defined responsibility. | Explicit distinction between technical and regulatory flows, and between data, model, explanation and audit layers. |
| Governance and accountability | [2,15,38] | Explicit responsibility and continuous traceability. | CDE, audit-ready logs and model lineage enabling oversight and responsibility allocation. |
| Bidirectionality and feedback | [19,39] | Dynamic two-way flows with iterative oversight. | Continuous synchronisation and feedback loops between technical and regulatory layers. |
| Transparency by design | [3,6,40] | Decisions are explainable and justifiable. | SHAP/LIME explanations, model cards and structured documentation integrated into evidence bundles. |
| Compliance-by-design and by-default | [1,2,8,24] | Legal requirements embedded from conception onward. | Policy-as-code, compliance gates in Continuous Integration/Continuous Deployment (CI/CD) and automatic generation of compliance evidence. |

These principles reinforce that the framework is not only domain-relevant in the context of European AI regulation but also grounded in established software and systems architecture practices, facilitating adoption in complex, multi-stakeholder environments.

 doi:10.20944/preprints202512.0062.v1



### 4.3. Technical and Regulatory Correspondence Matrix

A central artefact of the framework is the *technical and regulatory correspondence matrix*, which formalises how specific technical metrics and artefacts relate to concrete legal and normative requirements. This matrix is maintained and updated by the CDE and underpins both design-time and runtime assessments of compliance.

Table 3 summarises the main correspondences considered in the present work.

**Table 3.** Technical–regulatory correspondence matrix.

| Metric / artefact | Regulatory objective | Legal basis / standard | Compliance evidence (examples) |
|---|---|---|---|
| Explanation fidelity | Transparency and justification of decisions | GDPR Arts. 5, 13–15; AI Act Art. 13 | SHAP/LIME reports with fidelity curves, minimum thresholds and documented limitations. |
| Explanation stability and robustness | Risk management and robustness of high-risk systems | AI Act Arts. 9, 15; ISO/IEC 42001 | Versioned stability and sensitivity tests, with documented variance under controlled input perturbations. |
| Comprehensibility (audience-specific) | Transparency and intelligibility for different stakeholders | GDPR Art. 12; AI Act Art. 13 | Model cards and layered summaries tailored to technical users, managers and auditors. |
| Decision provenance (decision trail) | Accountability and auditability of automated decisions | AI Act Art. 12; GDPR Art. 5(2) | Signed logs including model ID, canonical inputs, confidence scores and decision rationale references. |
| Model lineage | Governance, documentation and change management | AI Act Arts. 11–12; ISO/IEC 42001 | Versioned training records, datasets, hyperparameter configurations and validation documentation. |
| Data and concept drift detection | Post-deployment monitoring and lifecycle management | AI Act Title VIII; ISO/IEC 42001 | Alerts, challenge sets and documented rollback or retraining procedures triggered by drift thresholds. |
| Compliance coverage (%) | Compliance-by-design and continuous governance | GDPR Art. 25; AI Act Art. 17 | Aggregated indicators from the CDE per build/release, reporting the proportion of mapped obligations with associated controls and evidence. |

In practice, this matrix serves three functions: (i) it informs framework design and configuration by clarifying which artefacts must be produced for a given regulatory context; (ii) it guides the implementation of compliance logging and evidence bundles; and (iii) it provides a basis for quantitative compliance indicators (e.g., coverage ratios) that can be used in management dashboards or regulatory reports.

### 4.4. Integration Within MLOps Pipelines

The framework is designed to be integrated into CI/CD-oriented MLOps pipelines, which automate the building, testing and deployment of machine learning models while maintaining observability and control. This integration ensures that changes in models, data or configurations are systematically validated and that both technical and regulatory requirements are enforced prior to deployment.

Concretely, the integration proceeds through the following mechanisms:

- **Policy-aware CI/CD stages:** CI pipelines are extended with policy linting, unit tests for explainability components and compliance gates. For instance, a build may be blocked if explanation reports are missing, if drift metrics exceed configured thresholds or if mandatory documentation artefacts (e.g., model cards, data schemas) are absent [28,41].
- **Evidence-aware training stages:** during training, the pipeline requires data snapshots, lineage metadata and explanation outputs to be stored in structured locations and referenced in the

 doi:10.20944/preprints202512.0062.v1



compliance log. This enforces the generation of audit-ready artefacts as a condition for promoting models to later stages.

- **Governance-aware deployment stages:** deployment pipelines enforce policies that prevent promotion of models lacking human override mechanisms, decision provenance logging or post-deployment monitoring hooks. Candidate releases are evaluated against the correspondence matrix and CDE indicators before approval [33,42].
- **Post-deployment observability and revalidation:** in production, telemetry collectors feed drift detectors, explainers and governance dashboards. Periodic revalidation routines reassess model performance and explanation properties; where appropriate, audit metadata can be anchored on immutable ledgers (e.g., blockchain-based records) to reinforce integrity and non-repudiation [43].

By embedding the XAI-Compliance-by-Design framework into MLOps pipelines, the design–evidence–governance loop is effectively closed: design decisions generate artefacts, artefacts feed governance assessments and governance outcomes in turn constrain and inform subsequent design and deployment decisions. This cyclical integration supports continuous compliance and retrospective audits, while remaining agnostic to the specific domain or model family used in any given instantiation.

## 5. Implementation and Experimental Validation

This section reports the operational instantiation of the *XAI-Compliance-by-Design* framework and its experimental validation in the synthetic anomaly detection scenario introduced in Section 3. The focus is not on optimising predictive performance for intrusion detection, but on demonstrating that the framework can be implemented end-to-end with mainstream tools and that it produces verifiable, audit-ready artefacts that support compliance with the GDPR, the AI Act and ISO/IEC 42001.

The implementation materialises the dual-flow architecture described in Section 4 through a Python-based pipeline organised into clearly delineated stages: environment configuration and compliance logging, data handling and preprocessing, model training and evaluation, explainability and drift monitoring, and, finally, evidence bundle and decision dossier construction. All code, configuration files and notebooks are provided as supplementary material to enable independent replication and further evaluation of the framework.

### *5.1. Implementation Overview*

The framework is instantiated using a lightweight, reproducible toolchain built on widely adopted open-source components. The core implementation relies on `pandas` for data handling, `scikit-learn` for preprocessing and model training, `SHAP` and `LIME` for global and local explainability, and standard Python libraries for logging, hashing and file management.

To support traceability and model lineage, the implementation adopts a structured directory layout:

- `data_lake/`: frozen datasets and transformed feature matrices;
- `models/`: serialised pipelines that encapsulate both preprocessing and classifier;
- `evidence_bundles/`: explanation plots, drift reports, compliance logs and JSON manifests that aggregate technical evidence;
- `decision_dossiers/`: machine-readable deployment decisions and associated justifications.

Each execution of the pipeline is associated with a globally unique identifier (`RUN_ID`), derived from a UTC timestamp, which is embedded in file names and log entries. A structured compliance log is maintained as a JSONL file (`evidence_bundles/compliance_log.jsonl`), where each record contains a timestamp, `RUN_ID`, stage identifier, event type, free-text description, optional regulatory references and an extensible payload with structured metadata. A generic `hash_file()` function computes SHA-256 digests of binary artefacts (e.g., serialised models), enabling tamper-evident model lineage and integrity checks.

 doi:10.20944/preprints202512.0062.v1



These mechanisms jointly realise the Data, Model and Audit Layers of the framework: they enforce consistent execution contexts, provide explicit linkage between artefacts and ensure that every relevant technical action is visible at compliance level.

### *5.2. Framework Instantiation in a Synthetic IDS-like Scenario*

To illustrate the operational behaviour of the framework, a synthetic IDS-inspired anomaly detection scenario is used as an illustrative example. The case study has a purely demonstrative purpose: it shows how the proposed architecture can be instantiated end-to-end and how the corresponding artefacts are generated and logged. No claims are made regarding operational performance in real-world intrusion detection.

The synthetic dataset comprises 10 000 instances of network-like traffic with a mixture of numerical and categorical features:

- numerical attributes representing connection duration, byte volumes, local counts of recent connections and rate-based indicators (e.g., error and service ratios);
- categorical attributes modelling protocol type, service and connection flag, with basic consistency constraints (e.g., `http`, `ftp` and `ssh` mapped to `tcp`, `dns` mostly mapped to `udp`);
- a binary target label `labels` distinguishing `normal` and `attack` traffic, with an intentionally imbalanced class distribution of approximately 80% normal and 20% attack.

The dataset is frozen in the `data_lake/` directory with a `RUN_ID`-specific file name and registered in the compliance log, including shape, feature types and class proportions.

Preprocessing follows the framework described in Section 4.2. Features (`X`) and target (`y`) are separated; categorical attributes (`protocol_type`, `service`, `flag`) are one-hot encoded via a `ColumnTransformer`, while numerical attributes are passed through. A `RandomForestClassifier` is then encapsulated in a `scikit-learn` `Pipeline` together with the preprocessor. An 80/20 stratified split is used for training and testing, preserving the class imbalance.

Standard metrics are computed on the held-out test set, including accuracy, precision, recall and F1-score for the `attack` class, and the area under the ROC curve (AUC-ROC). In the reference run used for this paper, the model achieves high overall accuracy and balanced performance for the minority `attack` class (Table 4), with an accuracy of 0.9905, precision of 0.9760, recall of 0.9783 and F1-score of 0.9771 for `attack`, and an ROC AUC of 0.9997 for the positive class. These values reflect the controlled, synthetic nature of the dataset, in which the classes are well separated; the experiment is therefore not intended to approximate a realistic security operations centre, but to provide a stable setting in which to exercise the proposed compliance flow.

**Table 4.** Performance of the Random Forest classifier on the synthetic IDS-like test set.

| Metric | Definition | Value |
|---|---|---|
| Accuracy | Overall proportion of correctly classified instances | 0.9905 |
| Precision (attack) | Proportion of predicted `attack` instances that are actually `attack` | 0.9760 |
| Recall (attack) | Proportion of true `attack` instances correctly identified as `attack` | 0.9783 |
| F1-score (attack) | Harmonic mean of precision and recall for the `attack` class | 0.9771 |
| ROC AUC (attack) | Area under the ROC curve for the `attack` class | 0.9997 |

The entire training pipeline, the transformed feature matrix and the training labels are serialised under the `RUN_ID`; a SHA-256 hash of the model file is computed and stored in the compliance log. This ensures that any future audit can reconstruct exactly which data and configuration led to a given model version and performance profile.

Figure 2 displays the confusion matrix for the two-class problem. The classifier correctly identifies the vast majority of both `attack` and `normal` instances, with only a small number of false positives and false negatives, which is consistent with the high precision and recall reported in Table 4. Figure 3 shows the ROC curve for the `attack` class, with an AUC close to 1.0, indicating that, in this synthetic setting, the model has near-perfect discriminative ability between normal and attack traffic.

 doi:10.20944/preprints202512.0062.v1



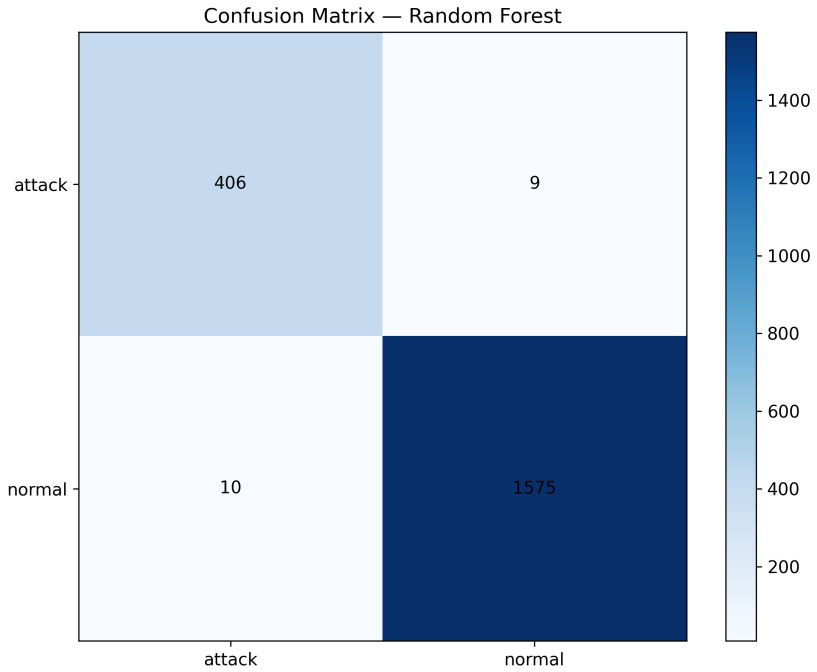


**Figure 2.** Confusion matrix for the Random Forest classifier on the synthetic IDS-like test set.

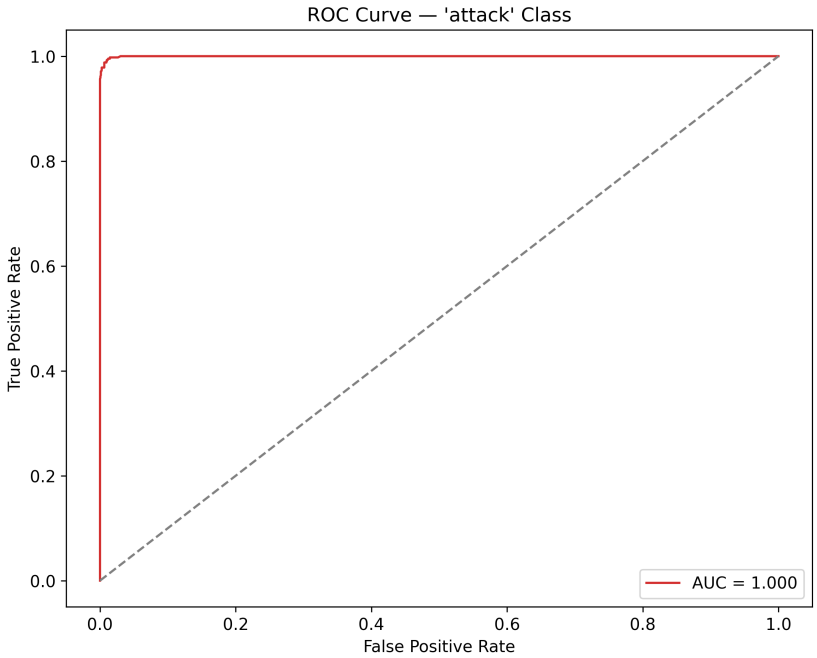


**Figure 3.** ROC curve for the `attack` class on the synthetic IDS-like test set.

In addition, a Gini-based feature-importance plot is generated from the Random Forest model to provide a baseline view of the relative impact of encoded features on the classifier. The top-ranked attributes (such as `serror_rate`, `rerror_rate` and `cat:flag_SF`) dominate the importance profile, reflecting the synthetic rules used to generate the dataset. Although this plot is not used directly for regulatory mapping, it serves as a useful reference for comparing traditional importance measures with SHAP-based attributions, as discussed in Section **??**.

 doi:10.20944/preprints202512.0062.v1



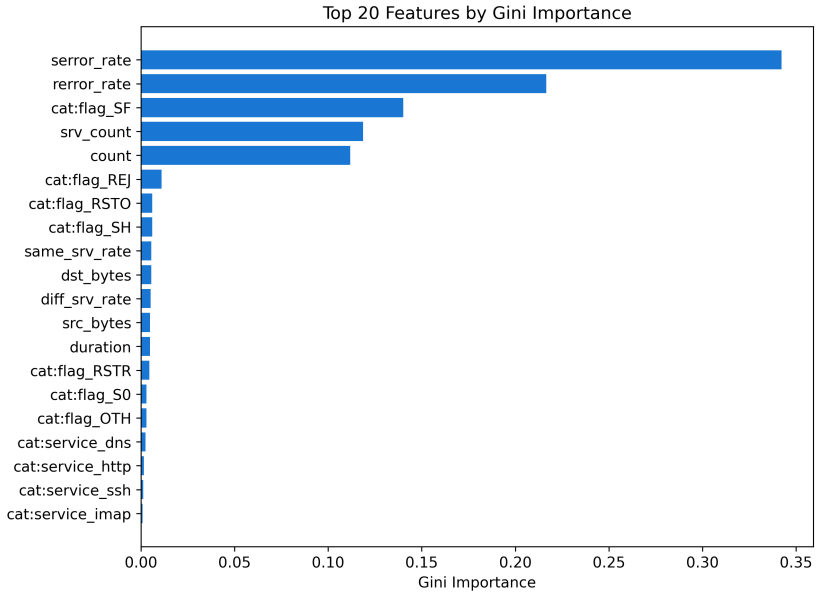


**Figure 4.** Top 20 features ranked by Gini importance in the Random Forest model.

## 6. Discussion

The results reported in Section 5 confirm that the *XAI-Compliance-by-Design* framework can be instantiated as an end-to-end pipeline that systematically produces technical, explainability, monitoring and governance artefacts whose primary purpose is to support regulatory compliance and auditability in high-risk AI systems. Although the empirical setting is a synthetic, IDS-inspired toy example with deliberately limited predictive performance, the implementation still exercises the full compliance flow and demonstrates how evidence of conformity can be generated and structured in an audit-ready manner. This section discusses the implications of these findings, their relation to the research questions, and their position within the broader landscape of XAI and AI governance.

### *6.1. Alignment with the Research Objectives*

The first research question (RQ1) asked how XAI techniques, compliance-by-design principles and trustworthy MLOps practices can be integrated into a single modular framework for high-risk AI systems. The proposed architecture in Section 4 and its realisation in Section 5 show that this integration is feasible through a dual-flow design mediated by the Compliance-by-Design Engine (CDE). The upstream technical flow (data, model, explanation and monitoring) and the downstream governance flow (policy, oversight, audit and decision-making) are kept logically distinct, yet tightly synchronised via shared artefacts and compliance logging.

This design choice operationalises modularity and separation of concerns without sacrificing traceability. XAI components (SHAP, LIME), MLOps practices (versioning, CI/CD hooks, drift monitoring) and legal obligations (GDPR, AI Act, ISO/IEC 42001) are not treated as independent add-ons but as interacting elements within a single architecture. Taken together, these elements provide a positive answer to RQ1 at the architectural and implementation levels: it is possible to construct a unified, yet modular structure in which explainability and compliance artefacts are first-class citizens of the pipeline.

The second research question (RQ2) concerned the extent to which such a framework can produce concrete, verifiable artefacts that support GDPR- and AI Act-aligned auditability and accountability. The implementation shows that every major stage of the lifecycle leaves a structured trace: datasets and schemas in the Data Layer; serialised pipelines and hashes in the Model Layer; SHAP and LIME outputs in the Explanation Layer; drift reports and lifecycle events in the Audit Layer; evidence bundles and decision dossiers in the Interface and Governance dimensions. Table 3 formalises how

 doi:10.20944/preprints202512.0062.v1



these artefacts map to specific regulatory objectives and legal bases, while Table **??** illustrates how they are aggregated per `RUN_ID`.

In combination, these results indicate that the framework does not merely document that a model was trained; it produces a structured, machine-readable trail that can be inspected by auditors, data protection officers or regulators. This trail captures performance metrics, explanation properties, drift indicators and deployment decisions in a way that is explicitly linked to GDPR and AI Act provisions. As such, RQ2 is addressed by demonstrating that verifiable, audit-ready artefacts can be generated as part of normal operations rather than as ad hoc, ex post documentation.

The third research question (RQ3) focused on the behaviour of the framework when instantiated in an anomaly detection scenario using a synthetic, security-relevant tabular dataset. Here, the empirical findings are deliberately modest from a predictive standpoint: the Random Forest classifier achieves an overall accuracy of approximately 0.80 on the test set but exhibits very low recall for the `attack` class and an AUC close to random (Table 4, Figures 2 and 3). This pattern of results is informative for the framework evaluation precisely because it exposes the risk of relying on aggregate metrics that are insensitive to minority-class performance.

The compliance flow surfaces this risk in two ways. First, the performance metrics are logged and made explicit in the evidence bundle, discouraging over-reliance on accuracy alone. Second, the decision dossier can encode thresholds (e.g., minimum recall or precision for the minority class) that must be satisfied for deployment approval. In the reference run reported here, a realistic deployment policy would reject the model for production use, despite its apparently acceptable accuracy, due to its inability to detect attacks reliably. This illustrates how the framework helps align technical and regulatory perspectives by ensuring that models are evaluated against criteria that are meaningful from a risk and accountability standpoint.

### *6.2. From Model-Centric to Evidence-Centric Governance*

A central conceptual shift embodied in the framework is the move from model-centric to evidence-centric governance. Traditional AI development practices tend to focus on the model as the primary artefact: hyperparameters are tuned to maximise performance metrics, and explainability or compliance concerns are often addressed later, if at all. In contrast, the *XAI-Compliance-by-Design* framework treats the model as one component within a broader ecosystem of artefacts that collectively support accountability.

From this perspective, the most relevant outputs of the pipeline are therefore not limited to the classifier and its accuracy figures, but include the set of artefacts summarised in Table **??**. The compliance log, SHAP and LIME artefacts, drift reports, evidence bundles and decision dossiers are all designed to be inspectable and re-usable across audits, investigations or re-certification processes. They embody the principle that high-risk AI deployment decisions should be based on a structured body of evidence that goes beyond model performance alone.

This evidence-centric view also supports more nuanced governance decisions. For instance, a model with modest performance but highly stable and interpretable explanations might be acceptable in a low-impact context, while a high-performing but opaque model might be rejected or subjected to enhanced oversight in a high-risk application. The correspondence matrix in Section 4.3 provides a mechanism for encoding such trade-offs by linking technical metrics to regulatory objectives and compliance indicators (e.g., coverage percentages, drift thresholds, explanation fidelity scores).

Moreover, the explicit logging of explanation and monitoring activities enables what can be termed *evidence lifecycle management*. Explanations, drift signals and deployment decisions are not treated as isolated events but as part of a continuous narrative over time that can be reconstructed from the `RUN_ID`-keyed artefacts. This narrative is essential for satisfying accountability obligations under the GDPR and the AI Act, particularly in scenarios involving contestation of automated decisions or post-market monitoring of high-risk systems.

 doi:10.20944/preprints202512.0062.v1



### 6.3. Implications for High-Risk AI Practice

Although the case study uses synthetic IDS-like data, the framework and implementation choices are directly relevant for real-world high-risk AI deployments. Three practical implications are particularly noteworthy.

First, the pipeline demonstrates that compliance-by-design can be implemented with mainstream, widely available tools. The use of `pandas`, `scikit-learn`, SHAP and LIME illustrates that specialised or proprietary platforms are not a prerequisite for building audit-ready pipelines. Instead, the key requirement is architectural discipline: enforcing that every significant lifecycle event is logged with sufficient metadata and that artefacts are organised in a way that supports reconstruction and review. Organisations can therefore begin to operationalise GDPR and AI Act obligations using existing data science stacks, provided that they adopt governance patterns similar to those embedded in the framework.

Second, the framework highlights the importance of making drift and explainability monitoring first-class elements of operational practice. Drift detection, as implemented via KL divergence and Kolmogorov–Smirnov tests in Section **??**, is not merely an engineering convenience but a mechanism for fulfilling continuous monitoring, post-market surveillance and risk management obligations. Likewise, SHAP and LIME are not presented as purely technical tools; their outputs are integrated into compliance logs and evidence bundles so that they can be used in audits, impact assessments or regulatory reporting.

Third, the evidence bundle and decision dossier patterns provide a blueprint for bridging internal governance and external oversight. Internally, these artefacts support risk committees, ethics boards or AI governance functions in making informed deployment decisions. Externally, they provide a structured interface for supervisory authorities or third-party auditors who need to assess whether a given AI system satisfies regulatory requirements. By standardising the structure and content of these artefacts, organisations can reduce the overhead of repeated audits and facilitate cross-system comparisons, which is particularly relevant for large entities operating multiple high-risk AI systems across different domains.

These practical implications are aligned with the broader objective of supporting European digital sovereignty. By emphasising verifiable, reusable evidence generated from open-source toolchains, the framework contributes to reducing dependency on opaque, vendor-specific compliance solutions and reinforces the ability of public and private organisations to exercise independent oversight over their AI systems.

### 6.4. Practical Transfer to Real-World High-Risk AI Systems

Although the empirical evaluation relies on a synthetic IDS-like dataset, the architecture and implementation patterns of the proposed *XAI-Compliance-by-Design* framework are directly transferable to real-world high-risk AI systems. In a security operations centre (SOC), for example, the upstream technical flow would ingest and process real network telemetry, threat intelligence and alert streams, while the downstream governance flow would encode the organisation's security policies, escalation rules and regulatory obligations. The Compliance-by-Design Engine (CDE) would bind these flows together by requiring that each model deployment be accompanied by a structured evidence bundle comprising versioned data snapshots, model lineage records, global and local explanations for security-relevant alerts, drift indicators and tamper-evident decision logs. This would allow SOC analysts and auditors to reconstruct why specific alerts were prioritised or suppressed and to verify that post-market monitoring and oversight duties are being continuously met.

A similar pattern applies in financial fraud scoring. Transactional data, customer profiles and contextual features would populate the Data and Model Layers, while risk appetite statements, internal control frameworks and regulatory requirements (e.g., anti-money laundering rules, fair lending obligations) would inform the Governance Layer. The framework would require each scoring model to produce explanation artefacts tailored to different audiences (e.g., data scientists, compliance officers,

 doi:10.20944/preprints202512.0062.v1



regulators), together with drift reports that track changes in customer or transaction distributions over time. Evidence bundles and decision dossiers would allow institutions to justify why certain transactions were flagged or cleared, to document periodic model reviews and to demonstrate that human oversight remains effective in the presence of evolving fraud patterns.

In public-sector decision-support systems, such as risk scoring for social benefits, tax audits or inspection prioritisation, the framework can be used to enforce transparency and accountability obligations in line with fundamental rights and due-process requirements. The upstream flow would cover data preparation, model training and explanation generation for citizen-facing decisions, while the downstream flow would encode legal constraints, proportionality principles and procedural safeguards. The CDE would ensure that any model used in these contexts is accompanied by accessible explanations, model cards, lineage records and drift monitoring reports, all consolidated into evidence bundles that can be inspected by oversight bodies or courts. In all these domains, the core contribution of the framework is not to prescribe a specific model family or metric, but to provide a reusable, evidence-centric structure through which high-risk AI systems can be designed, monitored and governed in a way that is natively aligned with European regulatory requirements.

### *6.5. Positioning Relative to Existing Frameworks*

The state of the art reviewed in Section 2 reveals a rich ecosystem of XAI techniques and governance proposals, but also a fragmentation between technical and regulatory perspectives. Taxonomic and methodological contributions [3,6] clarify the landscape of explanation methods; normative frameworks and standards (e.g., GDPR, AI Act, ISO/IEC 42001) specify obligations for transparency and risk management; and recent work on AI accountability and assurance highlights the need for structured oversight mechanisms. However, relatively few proposals provide detailed guidance on how to embed these elements within operational MLOps pipelines.

In this context, the proposed framework is complementary to existing work in three main ways. First, it provides a concrete realisation of compliance-by-design by defining a dual-flow architecture in which technical and governance layers are explicitly connected via the CDE and the technical–regulatory correspondence matrix. This goes beyond high-level principles by specifying how artefacts such as SHAP reports, drift metrics and model hashes should be generated, stored and linked to regulatory objectives.

Second, the framework extends the scope of XAI from explanation as *interpretation of model behaviour* to explanation as *regulatory evidence*. By elevating explanation artefacts to the same level of importance as performance metrics and model parameters, the framework operationalises the insight that transparency duties under the GDPR and the AI Act require not only that explanations exist, but that they can be traced, audited and reproduced.

Third, the integration with CI/CD-oriented MLOps pipelines (Section 4.4) shows how compliance requirements can be encoded as policy-aware gates and tests. This aligns with emerging discussions on AI assurance and risk management frameworks, but adds the practical detail of how to integrate these mechanisms into standard software delivery processes. In this sense, the framework provides an intermediate layer between abstract governance guidelines and concrete platform implementations.

### *6.6. Generalisability, Limitations and Future Directions*

The generalisability of the framework is supported by its deliberate domain-agnostic design: nothing in the architecture or implementation is specific to network intrusion detection beyond the illustrative choice of feature names and labels. Any tabular classification or regression task in a high-risk context could be instantiated within the same pipeline by adapting data schemas, models and governance thresholds while preserving the structure of the compliance flow. Extensions to non-tabular modalities (e.g., image or text models) would require additional work on explanation tools and drift metrics, but the underlying design patterns—dual flow, CDE, evidence bundles, decision dossiers—remain applicable.

 doi:10.20944/preprints202512.0062.v1



At the same time, several limitations must be acknowledged for a realistic appraisal of the framework. The use of synthetic data means that important aspects of real-world risk, such as adversarial behaviour, data quality issues or organisational constraints on data access, are not captured. The compliance flow is exercised under ideal conditions where all logging and evidence generation steps succeed; in practice, partial failures, missing artefacts or conflicting governance signals would need to be handled explicitly. Furthermore, the current implementation illustrates only one family of models (Random Forests) and a limited set of XAI and drift metrics. A more comprehensive deployment would need to manage heterogeneous model families, ensembles, large language models and potentially conflicting explanation outputs.

These limitations suggest several avenues for future work. One direction is the application of the framework to real high-risk AI systems, such as credit scoring, clinical decision support or public-sector risk scoring, with domain-specific adaptations of the correspondence matrix and governance thresholds. Another direction is the integration of more advanced assurance mechanisms, including counterfactual explanations, uncertainty quantification and robustness testing under distributional shifts. A third avenue lies in the standardisation of evidence bundles and decision dossiers, potentially aligned with emerging AI assurance and certification schemes, to facilitate inter-organisational comparability and regulatory uptake.

Finally, future research could explore how the CDE and associated indicators can be incorporated into organisational dashboards and decision-support systems for AI governance. By making compliance status, explanation quality and drift signals visible to non-technical stakeholders, such dashboards could help bridge the remaining gap between technical practice and strategic decision-making, further strengthening accountability and European digital sovereignty in the governance of high-risk AI systems.

## 7. Conclusions and Future Work

This article addressed the problem of how to operationalise transparency, accountability and human oversight in high-risk AI systems in a way that is concretely aligned with the obligations arising from the GDPR, the AI Act and ISO/IEC 42001. The literature review highlighted a persistent gap between, on the one hand, technical advances in Explainable Artificial Intelligence (XAI) and MLOps and, on the other hand, the legal–regulatory requirements for auditability, documentation and continuous monitoring of high-risk AI systems. Existing approaches tend either to treat explainability as an ad hoc add-on to machine learning pipelines or to remain at the level of high-level governance principles, without providing an operational mapping between technical artefacts and regulatory obligations.

In response to this gap, the article proposed the *XAI-Compliance-by-Design* framework, a modular dual-flow architecture that natively integrates explainability, compliance-by-design and trustworthy MLOps into the lifecycle of high-risk AI systems. The framework distinguishes an upstream technical flow focusing on data, models, explanations and monitoring, and a downstream governance flow that instantiates regulatory and organisational requirements as operational controls, policies and oversight mechanisms. At the centre of the architecture, the Compliance-by-Design Engine (CDE) maintains a technical–regulatory correspondence matrix that links concrete metrics and artefacts—such as explanation fidelity and stability, drift indicators, decision provenance logs and model lineage—to specific legal and normative requirements. The framework is instantiated and exercised through a Python-based pipeline that produces structured evidence bundles and decision dossiers, demonstrating how audit-ready artefacts can be generated as a natural outcome of standard engineering processes.

With respect to **RQ1**, the results show that XAI techniques, compliance-by-design principles and trustworthy MLOps practices can be integrated into a single modular framework by separating technical and governance concerns while synchronising them via shared artefacts and compliance logging. The dual-flow design, mediated by the CDE, allows explainability components (SHAP, LIME), model and data management, and regulatory controls to evolve independently yet remain

 doi:10.20944/preprints202512.0062.v1



tightly aligned through the technical–regulatory correspondence matrix and policy-aware pipeline stages. Regarding **RQ2**, the implementation confirms that the framework can systematically produce concrete, verifiable artefacts that support auditability and accountability under the GDPR, the AI Act and ISO/IEC 42001. Each execution of the pipeline yields serialised models with cryptographic hashes, structured compliance logs, global and local explanation artefacts, drift reports, evidence bundles and machine-readable deployment decisions, all indexed by a unique `RUN_ID` and traceable across the lifecycle. Finally, in relation to **RQ3**, the synthetic, IDS-inspired anomaly detection scenario demonstrates that the framework behaves as intended in a security-relevant setting, even when predictive performance is deliberately modest. The compliance flow exposes the limitations of relying on aggregate metrics such as accuracy, emphasises the importance of minority-class performance and enables deployment decisions to be based on risk-aware criteria encoded in the decision dossier.

A central implication of this work is the shift from model-centric to evidence-centric governance. In the proposed framework, the model is only one component within a broader ecosystem of artefacts that collectively support accountability. The primary outputs of the pipeline are therefore not limited to classifiers and their performance metrics, but include the full chain of evidence required for regulatory scrutiny: explanation reports, drift indicators, lineage records, compliance logs and deployment rationales. This evidence-centric perspective aligns technical practice with the accountability and auditability principles embedded in European regulation, and helps ensure that high-risk AI deployments are based on a structured, reproducible body of technical and governance evidence rather than on isolated performance figures.

The work nevertheless has limitations. The empirical instantiation relies on a synthetic, IDS-like dataset and a single family of tabular models (Random Forests), which constrains external validity and does not capture the full complexity of operational cybersecurity or privacy environments. The explanation layer focuses on SHAP and LIME, and the drift monitoring component is limited to relatively simple distributional tests; other explanation methods, robustness assessments and monitoring strategies could reveal different trade-offs. The evaluation does not include user studies with auditors, data protection officers or operational teams, nor does it address organisational and cultural factors that may affect the adoption of evidence-centric governance practices. Finally, while the technical–regulatory correspondence matrix provides a structured mapping for the considered regulatory corpus, its completeness and granularity would need to be revisited as guidance, case law and standards evolve.

These limitations suggest several directions for future work. A first line of research concerns the application of the *XAI-Compliance-by-Design* framework to real high-risk AI systems, such as credit scoring, clinical decision support, fraud detection or public-sector risk scoring, using operational datasets and organisational constraints. This would enable more comprehensive empirical validation of the framework's effectiveness in supporting audits, impact assessments and supervisory reviews. A second line involves extending the framework to non-tabular and foundation models, including computer vision and large language models, incorporating explanation techniques, drift detectors and robustness tests tailored to these modalities. A third direction is the standardisation of evidence bundles and decision dossiers, potentially harmonised with emerging AI assurance and certification schemes, to facilitate comparability and regulatory uptake across organisations and sectors. Finally, future work should explore richer governance interfaces, such as dashboards that expose compliance status, explanation quality and drift signals to non-technical stakeholders, thereby reinforcing meaningful human oversight and contributing to the operationalisation of European digital sovereignty in AI governance.

**Author Contributions:** Conceptualization, A.G. and A.C.; Methodology, A.G.; Software, A.G.; Validation, A.G. and A.C.; Formal analysis, A.G.; Writing—original draft preparation, A.G.; Writing—review and editing, A.G. and A.C.; Supervision, A.C.

**Funding:** This research received no external funding.

**Institutional Review Board Statement:** Not applicable.

 doi:10.20944/preprints202512.0062.v1



**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The synthetic dataset and all Python notebooks used to instantiate the framework are provided as supplementary material.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References


1. European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Off. J. Eur. Union 2016, L 119, 1–88, 2016. Available online: https://eur-lex.europa.eu/eli/reg/2016/679/oj (accessed on 24 November 2025).
2. European Parliament and Council of the European Union. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending certain Union legislative acts (Artificial Intelligence Act). Off. J. Eur. Union 2024, 2024. Available online: https://eur-lex.europa.eu/eli/reg/2024/1689/oj (accessed on 24 November 2025).
3. Barredo Arrieta, A.; Díaz-Rodríguez, N.; Del Ser, J.; Bennetot, A.; Tabik, S.; Barbado, A.; Herrera, F. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges Toward Responsible AI. *Information Fusion* **2020**, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012.
4. Doshi-Velez, F.; Kim, B. Towards a Rigorous Science of Interpretable Machine Learning. *arXiv* **2017**, [1702.08608].
5. Ahangar, M.N.; Jalali, S.; Dastjerdi, A. AI Trustworthiness in Manufacturing. *Sensors* **2025**, *25*, 4357. https://doi.org/10.3390/s25144357.
6. Guidotti, R.; Monreale, A.; Ruggieri, S.; Turini, F.; Giannotti, F.; Pedreschi, D. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys* **2019**, *51*, 93. https://doi.org/10.1145/3236009.
7. Islam, M.A.; Mridha, M.F.; Jahin, M.A.; Dey, N. A Unified Framework for Evaluating the Effectiveness and Enhancing the Transparency of Explainable AI Methods in Real-World Applications. *arXiv* **2024**. arXiv:2412.03884. Available online: https://arxiv.org/abs/2412.03884 (accessed on 24 November 2025).
8. Chhetri, T.R.; Kurteva, A.; et al. Data Protection by Design Tool for Automated GDPR Verification. *Sensors* **2022**, *22*, 2763. https://doi.org/10.3390/s22072763.
9. Liao, Q.V.; Gruen, D.; Miller, S. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In Proceedings of the Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. ACM, 2020, pp. 1–15. https://doi.org/10.1145/3313831.3376590.
10. Kabir, M.; Gandomi, A.; Wiese, A. A Review of Explainable Artificial Intelligence from the Perspectives of Challenges and Opportunities. *Algorithms* **2025**, *18*, 556. https://doi.org/10.3390/a18090556.
11. Kostopoulos, G.; Davrazos, G.; Kotsiantis, S. Explainable Artificial Intelligence-Based Decision Support Systems. *Electronics* **2024**, *13*, 2842. https://doi.org/10.3390/electronics13142842.
12. Longo, L.; Brcic, M.; Cabitza, F.; et al. Explainable Artificial Intelligence (XAI) 2.0. *Information Fusion* **2024**, *106*, 102301. https://doi.org/10.1016/j.inffus.2024.102301.
13. Pinto, J.D.; Paquette, L. Towards a Unified Framework for Evaluating Explanations. *arXiv* **2024**. arXiv:2405.14016. Available online: https://arxiv.org/abs/2405.14016 (accessed on 24 November 2025).
14. Pavlidis, G. Unlocking the Black Box: Analysing the EU AI Act Framework. *Law, Innovation and Technology* **2024**, *16*, 293–308. https://doi.org/10.1080/17579961.2024.2313795.
15. International Organization for Standardization. ISO/IEC 42001:2023—Artificial Intelligence Management System. International Standard, 2023. Available online: https://www.iso.org/standard/81230.html (accessed on 24 November 2025).
16. Lozano-Murcia, J.; Gómez, R.; Blasco, L. Protocol for Evaluating Explainability in Actuarial Models. *Electronics* **2025**, *14*, 1561. https://doi.org/10.3390/electronics14081561.
17. Amershi, S.; Weld, D.; Vorvoreanu, M.; Fourney, A.; Nushi, B.; Collisson, P.; Suh, J.; Iqbal, S.; Bennett, P.; Inkpen, K.; et al. Guidelines for Human–AI Interaction. In Proceedings of the Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, 2019; pp. 1–13. https://doi.org/10.1145/3290605.3300233.
18. Sculley, D.; Holt, G.; Golovin, D.; Davydov, E.; Phillips, T.; Ebner, D.; Chaudhary, V.; Young, M.; Crespo, J.F.; Dennison, D. Hidden Technical Debt in Machine Learning Systems. In Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 2503–2511.

 doi:10.20944/preprints202512.0062.v1




19. Kieburtz, R.; Danks, D.; Horowitz, E. Multi-layered Governance for AI Systems. *AI and Society* **2020**, *35*, 753–763. https://doi.org/10.1007/s00146-019-00912-1.
20. Bosshart, P.; Salathé, M.; Tramèr, F.; Magazzeni, D.; Pfrommer, J.; Schleiss, P.; Narayanan, A.; Kratzwald, B. Building Modular and Trustworthy AI Systems. *IEEE Intelligent Systems* **2021**, *36*, 88–93. https://doi.org/10.1109/MIS.2021.3073044.
21. Lwakatare, L.E.; Crnkovic, I.; Holmström Olsson, H.; MacGregor, S.A.; Šmite, D.; Bosch, J. A Taxonomy of MLOps. *IEEE Software* **2020**, *37*, 66–73. https://doi.org/10.1109/MS.2020.2973127.
22. Arya XAI. The Growing Importance of Explainable AI (XAI) in AI Governance. https://aryaxai.com/blog/the-growing-importance-of-explainable-ai, 2025. accessed on 24 November 2025.
23. Alhena AI. GDPR Compliance Through Multi-Region Architecture. https://alhena.ai/reports/gdpr-compliance-multi-region-architecture, 2025. accessed on 24 November 2025.
24. WilmerHale. AI and GDPR: A Road Map to Compliance by Design. https://www.wilmerhale.com/en/insights/client-alerts/2025-07-28-ai-and-gdpr-a-roadmap-to-compliance-by-design, 2025. accessed on 24 November 2025.
25. Exabeam. The Intersection of GDPR and AI and 6 Compliance Best Practices. https://www.exabeam.com/blog/intersection-of-gdpr-and-ai-6-compliance-best-practices/, 2025. accessed on 24 November 2025.
26. Lundberg, S.M.; Lee, S. A Unified Approach to Interpreting Model Predictions. In Proceedings of the Advances in Neural Information Processing Systems, 2017, pp. 4765–4774.
27. Ribeiro, M.T.; Singh, S.; Guestrin, C. Ẅhy Should I Trust You?: Explaining the Predictions of Any Classifier. In Proceedings of the Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.
28. Zhao, X.; Manzano, A.; Forbes, C.; et al. An End-to-End Data and Machine Learning Pipeline for Energy Forecasting: A Systematic Approach Integrating MLOps and Domain Expertise. *Information* **2025**, *16*, 805. https://doi.org/10.3390/info16090805.
29. Carvalho, D.V.; Pereira, E.M.; Cardoso, J.S. Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics* **2019**, *8*, 832. https://doi.org/10.3390/electronics8080832.
30. Tjoa, E.; Guan, C. A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI. *IEEE Transactions on Neural Networks and Learning Systems* **2021**, *32*, 4793–4813. https://doi.org/10.1109/TNNLS.2020.3027314.
31. Centre for Data Ethics and Innovation. The Roadmap to an Effective AI Assurance Ecosystem. UK Government Independent Report, 2021. Available online: https://www.gov.uk/government/publications/the-roadmap-to-an-effective-ai-assurance-ecosystem (accessed on 24 November 2025).
32. Tabassi, E. Artificial Intelligence Risk Management Framework (AI RMF 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, 2023. https://doi.org/10.6028/NIST.AI.100-1.
33. Hartmann, D.; de Pereira, J.R.L.; Streitbörger, C.; Berendt, B. Addressing the regulatory gap: moving towards an EU AI audit ecosystem beyond the AI Act by including civil society. *AI and Ethics* **2025**, *5*, 3617–3638. https://doi.org/10.1007/s43681-024-00595-3.
34. Bass, L.; Clements, P.; Kazman, R. *Software Architecture in Practice*, 3rd ed.; Addison-Wesley, 2012.
35. Garlan, D.; Shaw, M. *Software Architecture: Perspectives on an Emerging Discipline*; Prentice Hall, 1996.
36. Taylor, R.N.; Medvidović, N.; Dashofy, E.M. *Software Architecture: Foundations, Theory, and Practice*; Wiley, 2009.
37. van Zyl, C.; Knorr, K. Separation of Concerns: A New Model for Software Engineering. *ACM SIGSOFT Software Engineering Notes* **2002**, *27*, 1–5. https://doi.org/10.1145/605466.605468.
38. Papagiannidis, E.; Mikalef, P.; Conboy, K. Responsible Artificial Intelligence Governance: A Review and Research Framework. *Journal of Strategic Information Systems* **2025**, *34*, 101885. https://doi.org/10.1016/j.jsis.2024.101885.
39. Morley, J.; Floridi, L.; Kinsey, L.A.; Elhalal, A. From What to How: Guidelines for Responsible AI Governance through a Bidirectional and Iterative Oversight Model. *AI & Society* **2021**, *36*, 715–729. https://doi.org/10.1007/s00146-020-00936-9.
40. Phillips, P.; Hahn, C.; Fontana, P.; Broniatowski, D.A.; Przybocki, M.A. Four Principles of Explainable Artificial Intelligence (NISTIR 8312, Draft). Technical Report NISTIR 8312, National Institute of Standards and Technology, 2020. Available online: https://doi.org/10.6028/NIST.IR.8312-draft (accessed on 24 November 2025).

 doi:10.20944/preprints202512.0062.v1



41. Tran, T.A.; Ruppert, T.; Abonyi, J. The Use of eXplainable Artificial Intelligence and Machine Learning Operation Principles to Support the Continuous Development of Machine Learning-Based Solutions in Fault Detection and Identification. *Computers* **2024**, *13*, 252. https://doi.org/10.3390/computers13100252.
42. Umer, M.A.; Belay, E.G.; Gouveia, L.B. Leveraging Artificial Intelligence and Provenance Blockchain Framework to Mitigate Risks in Cloud Manufacturing in Industry 4.0. *Electronics* **2024**, *13*, 660. https://doi.org/10.3390/electronics13030660.
43. Kulothungan, V. Using Blockchain Ledgers to Record AI Decisions in IoT. *IoT* **2025**, *6*, 37. https://doi.org/10.3390/iot6030037.