National College of Ireland

Design of an XAI-Enhanced Machine Learning Audit Model for Ethical, Technical, and Regulatory Compliance in the European Financial Sector

MSc Research Project
MSc In Fintech

Jenyffer Noelia Mazat Ixcol
Student ID: 24112551

School of Computing
National College of Ireland

Supervisor: Victor del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Jenyffer Noelia Mazat Ixcol		
Student ID:	24112551		
Programme	MSc In Financial Technology	**Year:**	2024 -2025
Module:	MSc Research Project		
Supervisor:	Victor del Rosal		
Submission Due Date:	11th of August 2025		
Project Title:	Design of an XAI-Enhanced Machine Learning Audit Model for Ethical, Technical, and Regulatory Compliance in the European Financial Sector.		
Word Count:	6195 Words – 20 pages excluding references.		

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Jenyffer Mazat
Date:	11th of August 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

AI Acknowledgement Supplement

MSc Research Project (MSCFTD1)

Thesis

Your Name/Student Number	Course	Date
Jenyffer Noelia Mazat / x24112551	**MSCFTD1**	11th of August 2025

This section is a supplement to the main assignment, to be used if AI was used in any capacity in the creation of your assignment; if you have queries about how to do this, please contact your lecturer. For an example of how to fill these sections out, please click here.

AI Acknowledgement

This section acknowledges the AI tools that were utilized in the process of completing this assignment.

Tool Name	Brief Description	Link to tool
ChatGPT	Brainstorming research, and helping to improve the structure, readability, and functionality of my existing code.	https://chat.openai.com/
Grammarly Go	Grammar and punctuation checks	https://www.grammarly.com/grammarlygo
Mermaid	For organizing a diagram of a multi-process AI structure.	https://www.mermaidchart.com/

Description of AI Usage

This section provides a more detailed description of how the AI tools were used in the assignment. It includes information about the prompts given to the AI tool, the responses received, and how these responses were utilized or modified in the assignment. **One table should be used for each tool used**.

ChatGPT	
1. ChatGPT was used to brainstorm ideas and refine the initial concept for my project based on the research papers and gaps identified. 2. It also helped to enhance the structure, clarity, and overall functionality of my existing code, ensuring better organization, improved readability, and optimized performance for seamless execution and maintenance. 3. ChatGPT was utilized to guide me in finding the best way to log sessions, given its novelty for my own purposes.	
1. Sample prompts: "Can you help me brainstorm potential approaches and refine my initial concept for an AI framework for auditing financial institutions? I have identified the following gaps:" Limited integration between AI-based anomaly detection and compliance rule-checking. Lack of standardized data pipelines for multi-source audit information. Minimal research on explainable AI models tailored for audit reporting. Insufficient real-time processing capabilities for continuous auditing.	1. Sample response from ChatGpt: "awesome—let's turn those gaps into concrete options you can pick from, plus a tight MVP you can ship. North-star A continuous audit platform that fuses anomaly ML + rules, runs on standardized pipelines, explains itself like an auditor, and works in (near) real-time. 1) How to integrate ML anomalies with compliance rules Pick an integration pattern (you can mix) A. Rules-as-features: compile policy thresholds/windows into features used by ML (e.g., "rolling_24h_amt_vs_limit"). B. ML-as-policy-advisor: anomalies propose threshold/cohort tweaks; analysts accept → rules repo updates.

Few case studies validating AI frameworks in real-world audit scenarios.	C. Decision fusion (recommended): rules engine + ML both score → policy layer reconciles outcomes with explicit logic (e.g., consensus, override, or "novel only"). D. Constraint-aware training: encode "never miss rule-critical scenarios" as hard constraints or penalties in the loss."
2. Sample Prompt: "What is the best way to integrate logging sessions in a ML auditing Framework?"	2. Sample response: " Great question. In an ML auditing framework, "logging sessions" should feel like case files: every action and decision is captured, linked, and reproducible. Here's a tight blueprint you can lift into your stack. What a "logging session" should mean • A bounded context that ties together: • who/what/when (actor, component, time), • the data & model state (inputs, features, model/rule versions), • the decision (scores, rules fired, thresholds), • the explanation & evidence (attributions, counterfactuals, query hashes), • and the lineage (where the data came from, transformations)."
3. Sample prompt: "What is the difference between audit log and logging session?" a. Scaling and Normalization: Could you suggest efficient ways to prepare the merged dataset for machine learning, including scaling or normalization techniques?	3. Sample response: "Here's the short version: Audit log → a record of events (often system-wide) kept for accountability, compliance, or forensic purposes. It's a timeline of what happened. Logging session → a structured, bounded capture of everything related to one end-to-end process or analysis run (e.g., a model evaluation for a single transaction). It's a container for a set of related logs."
1. Sample prompt: "I need the code to build an diagram in Mermaid that follows the following steps: ✅ Block 1 Dataset and sampling ✅ Block 2 Column cleaning + validation ✅ Block 3Target variable construction ✅ Block 4 Preprocessing and regulatory encoding ✅ Block 5: Exploratory Data Analysis (EDA) ✅ Block 6 Base ML model training ✅ Block 7Technical validation and robustness ✅ Block 8Global and local explainability ✅ Block 9 Fairness audit ✅ Block 10 Legal compliance mapping ✅ Block 11 Final technical report"	4. Python Snippet output.

Grammarly Go	
Grammarly Go was used throughout the writing process for checking grammar, punctuation, and readability of the text.	
NA	NA

EVIDENCE OF AI USAGE

This section includes evidence of significant prompts and responses used or generated through the AI tool. It should provide a clear understanding of the extent to which the AI tool was used in the assignment. Evidence may be attached via screenshots or text.

Additional Evidence A:

Audit Log template:

```python
import json, time

def audit_log(event_type, **fields):
    log_entry = {
        "timestamp": time.strftime("%Y-%m-%dT%H:%M:%SZ", time.gmtime()),
        "event_type": event_type,
        **fields
    }
    print(json.dumps(log_entry))

# Example usage
audit_log("rule.update",
          actor_id="analyst_002",
          rule_id="AML-3",
          old_threshold=5000,
          new_threshold=4500)

audit_log("model.infer",
          actor_id="svc_anomaly",
          model_id="fraud_v1.4",
          score=0.92,
          transaction_id="txn_98231")
```

Log event:

```python
log_event(
    block="4.2",
    component="Rule Evaluation",
    action_performed="Evaluated AML-3 rule on transaction ID txn_98231",
    outcome="Breach detected: amount €5,550 exceeds €5,000 threshold",
    regulatory_compliance="Yes",
    referenced_regulations="EU AMLD5; FATF Recommendation 10; CBI AML Guidelines",
    evidence_file="txn_98231.json, rule_AML-3_v2.1.yaml",
    detected_risk="Potential money laundering",
    risk_level="High",
    recommendation="Escalate case to compliance team for enhanced due diligence",
    auditor_comment="Rule correctly identified breach; evidence file is complete and linked to rel
)
```

Design of an XAI-Enhanced Machine Learning Audit Model for Ethical, Technical, and Regulatory Compliance in the European Financial Sector

Jenyffer Noelia Mazat Ixcol
Student ID: 24112551

Abstract

The European financial sector is increasingly relying on Artificial Intelligence decision-making, which requires rigorous governance systems and transparent fair practices. The EU Artificial Intelligence Act, GDPR, DORA, EBA Guidelines and CBI expectations have established strict requirements for high-risk AI systems but provide insufficient operational direction resulting in a gap between regulatory expectations and institutional implementation.

This research develops and validates an AI Audit framework that upholds ethical principles, explainability methods (SHAP, LIME, EBM-native explainability) and EU–Irish regulatory requirements, by developing a twelve-pillar system with Disparate Impact, Statistical Parity Difference, Equal Opportunity Difference, fairness assessments, technical validation and automated compliance mapping into a unified auditable workflow.

The Lending Club data is used in a credit-scoring simulation, to generate auditable evidence which follows regulatory requirements for supervisory assessment, results show high predictive accuracy (ROC-AUC > 0.99), complete fairness threshold compliance (DI ≥ 0.80) and the ability to detect operational gaps, especially in Ethical Alignment, Explainability and Legal Basis for Processing from self-reported to evidence-based compliance. The framework enables financial institutions, consultancy firms and regulators to create standardized audit artefacts through a repeatable process which strengthens trust in high-risk AI systems while closing the regulatory gap.

1. Introduction

Financial institutions are increasingly relying on Artificial Intelligence (AI) to enhance operational efficiency; however, they are facing multiple regulatory, ethical, and governance issues. *The Artificial Intelligence Act (AI Act; European Parliament & Council, 2024), the General Data Protection Regulation (GDPR; European Parliament & Council, 2016), the Digital Operational Resilience Act (DORA; European Parliament & Council, 2022), the Loan Origination and Monitoring Guidelines (LOM; EBA, 2020), and the Central Bank of Ireland governance requirements (CBI, 2021)* establish strict obligations for transparency, fairness, human oversight, resilience, and accountability, but they do not specify operational procedures or testing protocols or standardized audit evidence formats.

The absence of procedural guidance has created a gap between regulatory requirements and their actual implementation. The regulations specify important results including fairness, explainability, and traceability, but they fail to provide clear guidelines on implementing these principles into daily AI model operations. Technical tools like SHAP, LIME and Fairlearn enable strong interpretability and fairness evaluation, but their integration into standardized processes which generate supervisory review evidence is uncommon. Current research alongside supervisory communications focus on independent aspects of AI governance without creating an end-to-end compliance framework for the field.

This research addresses the existing gap by creating and validating an AI Regulatory Audit Framework which financial institutions under regulation can implement. A single standardized process through this framework produces dependable evidence that combines technical validation results with fairness assessment findings, together with regulatory compliance mapping.

The research investigates the following question:

"Can a technical, replicable model be designed to audit the use of AI in financial institutions in accordance with ethical principles, explainability, and the current regulations in Ireland and the EU?"

The study begins by examining EU AI Act provisions together with Irish financial regulations that apply to AI system deployment. The research determines which ethical principles and explainability requirements are most important for financial AI model evaluation while extending this assessment to include legal and technical aspects. The research uses Python to create a technical verification model which combines explainability (XAI) methods with fairness metrics and traceability mechanisms to perform automated evidence-based auditing procedures. The model undergoes validation through a simulated credit-scoring case study using public data to evaluate its operational effectiveness. The thesis presents a repeatable method for financial institutions, consultants, and regulators to ensure ongoing AI compliance.

The developed AI Regulatory Audit Framework provides complete technical implementation of twelve compliance pillars through auditable tests alongside structured logging and automated compliance mapping. The framework transforms complex EU and Irish regulatory requirements into practical audit evidence which organizations can apply across different financial institutions.

The dissertation structure includes section 2, which reviews the regulatory, ethical, and technical background before presenting twelve compliance pillars. Section 3 on research methodology explains steps for dataset preparation, model selection, and regulatory-to-technical mapping. The framework development process appears in sections 4 and 5, while section 6 presents performance evaluation, interpretability assessment, fairness analysis, and compliance. Finally, section 7 provides findings, research boundaries, and future directions for academic and professional implementation.

2. Related Work

The thesis integrates AI auditing, EU/IE financial regulations with XAI, and algorithmic fairness foundations through this literature review. The review uses regulatory documents and academic literature to create a critical synthesis of prior approaches and presents their strengths, weaknesses, methodological differences, and the existing gaps that drive the development of a unified audit system.

2.1 EU financial services sector standards for AI implementation

The Artificial Intelligence Act (EU AI Act) classifies creditworthiness determination AI applications and credit score generation as high-risk requiring risk management with data governance, technical documentation, transparency, human oversight, robustness, and post-market surveillance. Financial institutions face an implementation gap because the Act does not specify governance procedures and testing protocols, and audit evidence requirements (European Parliament & Council, 2024).

The General Data Protection Regulation (GDPR) provides essential guidelines about data processing, disclosure requirements and the protection of data subject rights when AI-based decisions are made. The disclosure obligations (Articles 13-15) and the safeguard requirements of Article 22 form the basis for data subject rights in AI-based decision-making systems. Studies by Nannini et al. (2023), Gyevnar et al. (2023), and the European Parliament & Council (2016) indicate that legal explanation requirements create tensions with practical explainability needs that require documentation standards for provenance and model reporting to ensure testability of these obligations.

The Digital Operational Resilience Act (DORA) establishes ICT risk management standards while requiring incident reporting along with resilience testing and third-party oversight throughout the financial sector. The monitoring of data pipelines and retraining processes, and change management operations fall under AI risk management requirements. AI lifecycle operations will need to prove continuous control and traceability starting from January 2025 according to European Parliament & Council (2022).

The EBA Guidelines on Loan Origination and Monitoring (EBA/GL/2020/06) stress governance, creditworthiness assessment, data quality, documentation, validation, and borrower-focused outcomes in ML-based lending. The lack of specifications about explainability metrics, fairness measures, and automated audit tools in the guidelines creates challenges for turning them into testable criteria (EBA, 2020).

The Central Bank of Ireland (CBI) requires detailed oversight, documentation, and risk management for AI components in third-party agreements across all sectors (CBI, 2021). The Regulatory & Supervisory Outlook 2025 from the Central Bank of Ireland lists operational resilience and AI governance, human oversight, and consumer protection as its main priorities

but does not provide specific technical criteria for XAI and fairness testing (CBI, 2025; Kincaid, 2024).

EU and IE regulations share common objectives regarding transparency, fairness, oversight, robustness, and resilience, but they do not provide operational guidelines for creating audit-grade evidence throughout explainability, fairness, documentation, and resilience in AI decisions. The governance duties of firms and board accountability receive reinforcement from supervisory communications, but the need for technical measures and evidence formats remains with the need for a unified audit framework (ESMA, 2024; EIOPA, 2021, 2025).

2.2 Explainability in high stakes decisioning: methods and limits

XAI research provides methods to explain AI decisions after they are made and methods that allow for direct interpretation of AI models. The SHAP technique merges game theory with additive feature attribution to deliver both local accuracy and consistency guarantees (Lundberg & Lee, 2017; Hermosilla et al., 2025) while LIME provides surrogate explanations across different model classes (Ribeiro et al., 2016; Khan, 2025). Explainable Boosting Machines (EBMs) implement generalized additive models with modern boosting and interaction detection, which provide strong global interpretability while maintaining competitive accuracy performance (Mathew, 2025). The literature stresses that interpretability evaluation must be conducted through strict assessment methods while ensuring alignment with stakeholder objectives and regulatory requirements (Doshi-Velez & Kim, 2017; Kim, 2024).

Comparative view: Post-hoc XAI offers adjustable diagnostic capabilities for black-box models but produces unstable or misleading attributions during distribution shifts or when features are correlated; inherently interpretable models (e.g., EBMs) sacrifice raw accuracy for transparency and editability, which becomes beneficial for governance and change control purposes. The literature advocates for an investment strategy which uses interpretable models when possible, with robust post-hoc tools and documentation processes to fulfil explainability obligations.

2.3 Algorithmic fairness: metrics, trade-offs, and sector nuances

Metrics and definitions. Fairness measurement in the field takes the form of different criteria such as error-rate parity (e.g., equality of opportunity), independence (e.g., statistical parity), and calibration. Seminal works like Hardt, Price, & Srebro (2016) establish equality of opportunity definitions and Kleinberg et al. (2017) demonstrate that multiple fairness criteria can only be satisfied in some cases. Corbett-Davies and Goel (2018, 2023) examine how metric-driven fairness falls short and support evaluation methods which link to specific policies.

The fairness framework in regulated finance receives expansion through recent studies which include Nwafor et al. (2024) developing an explainable hybrid credit scoring model and Ferrara

(2023) analyzing bias origins and reduction methods; Casas et al. (2024) assessing fairness metrics in credit risk auditing; Kim et al. (2023) studying regulatory compliance between fairness and explainability; and Calegari (2023) delivering an integrated auditing framework for financial systems. The sector now implements solutions which unite technical metrics with governance structures to fulfil ethical and regulatory standards according to these works.

The Fairlearn toolkit (Weerts et al., 2023; Bird et al., 2020) allows users to detect group disparities, however, industry research demonstrates that only tools cannot achieve fairness because it needs socio-technical integration with defined thresholds and documentation protocols, and human supervision. The Loan Origination and Monitoring Guidelines (LOM; EBA, 2020) and ESMA/EIOPA regulatory bodies require evidence-based fairness documentation which supports approaches that combine distributional effects with regulatory compliance and consumer protection and profitability instead of using only parity metrics.

2.4 Auditing and documentation practices

The modern literature demonstrates a need for continuous auditing that merges technological tests with governance documentation practices. Raji et al. (2020) introduce an internal algorithmic auditing system which creates cumulative evidence artefacts through the entire lifecycle process. Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2021) improve transparency and reproducibility, yet require integration into ML pipelines (versioning, logging, and review) to fulfil regulatory standards. Supervisory communications (e.g., ESMA, 2024; EIOPA, 2021, 2025) emphasize board accountability together with risk management and proportionality, but do not eliminate the necessity for concrete audit procedures, the gap between governance principles and operational testing/reporting persists.

2.5 Critical comparison and limitations of prior work

The research reveals three common obstacles that run across all studied sources consulted:

a) **Fragmentation between legal principles and technical practice:** The regulatory framework sets performance criteria (explainability, fairness, oversight), however, it does not define testing plans or measurement thresholds, or documentation requirements. Toolkits measure but rarely align outputs to regulatory reporting formats or audits at scale. This creates a compliance translation gap in finance. (CBI, 2021; EBA, 2020; EIOPA, 2021, 2025; ESMA, 2024; European Parliament & Council, 2016, 2022, 2024).
b) **Metric-level focus without lifecycle integration:** Most academic research focuses on enhancing fairness and XAI metrics but there is limited work which demonstrates verifiable repeatable procedures linking measurements to financial supervisory demands for documentation, logging, change control, and resilience. (Hardt et al., 2016; Kleinberg et al., 2017; Weerts et al., 2023; Mitchell et al., 2019).

c) **Limited empirical validation in regulated financial contexts:** Public datasets analysed under regulatory constraints in finance-related case studies are scarce, which limits both regulatory learning and comparison capabilities. Supervisory sources maintain principle-based approaches without technical examples which forces firms to build their own evidence frameworks. (EBA, 2020; ESMA, 2024; EIOPA, 2021, 2025).

The field provides strong principles (EU AI Act, GDPR, DORA, sectoral guidance) and mature techniques (XAI, fairness metrics, documentation patterns). However, the current literature lacks functional auditing methods. Why are prior solutions not enough? Some governance documents are not measurable tests; also, technical papers rarely align outputs to regulatory reporting or lifecycle documentation. Toolkits enable the measurement of fairness and explainability but do not fulfil the requirements of oversight and traceability and resilience by themselves.

The thesis addresses existing gaps through its modular multi-pillar audit framework which integrates explainability with fairness, documentation, oversight and resilience into evidence-based workflows that align with EU AI Act and GDPR and DORA and EBA guidelines and CBI requirements developing a functional connection between academic methodologies and supervisory achievement.

3. Research Methodology

This research will combine predictive modelling techniques with regulatory auditing through an experimental case-study method to evaluate AI systems in financial applications. The methodological structure will follow a two-phase approach, the first phase for developing reproducible baseline models through credit scoring applications followed by the second phase where the AI Regulatory Audit Framework implements its main research contribution through a multi-pillar evidence-based system that meets EU and Irish regulatory standards.

The current design follows previous studies (Doshi-Velez & Kim, 2017; Raji et al., 2020) which demonstrate that technical evaluation approaches with governance frameworks enhance high-risk sector AI accountability. The audit framework receives its normative base from EU AI Act, GDPR, DORA, EBA, and CBI governance principles. The five regulations expectations support twelve interconnected pillars for regulated AI compliance which include (1) explainability; (2) fairness/bias; (3) traceability & logging; (4) human oversight; (5) auditability & reporting; (6) risk & robustness; (7) ethical alignment; (8) data minimization; (9) lawful basis; (10) resilience & retraining readiness; (11) impact assessment; and (12) documentation & record-keeping. Studies validate each pillar independently but few research papers create a unified auditing workflow that combines all pillars for financial AI applications. A unified auditing method which assesses multiple pillars simultaneously demonstrates the need for an integrated framework. (European Parliament & Council, 2016, 2022, 2024; EBA, 2020; CBI, 2021, 2025).

3.1 Data requirements, acquisition, fairness and governance

To simulate a ML credit risk scoring this research uses the Lending Club public loan data (2M+ and 151 variables) guaranteeing the exclusion of personal data. In line with GDPR Article 5 on data minimization a sample of 200,000 records is going to be extracted through stratified sampling with random selection while maintaining the target distribution by converting loan_status into a binary default indicator.

Execution environment and tooling: procedures will work in a Python environment in Google Colab with Google Drive persistence, employing libraries such as pandas, numpy, scikit-learn, interpret (EBM), shap, lime, fairlearn, matplotlib, seaborn, dill/joblib for model export and to ensure traceability all outputs or resources are going to be saved on Google Drive directory to comply with the documentation and record-keeping obligations (DORA, EU AI Act) by maintaining an immutable, versioned archive of evidence suitable for audit purposes.

Preprocessing data: removal of irrelevant/high-missingness, splitting: stratified 80/20 train–test with fixed seed (random_state=42) is needed to preserve class balance, and subsequent preprocessing within the training set including encoding of categorical variables via label/one-hot encoding, for LR median imputation, using SimpleImputer(strategy="median").

3.2 Phase 1 – Model selection

The first phase of baseline model preparation consists of acquiring data, preprocessing features and encoding them while training representative models that include Logistic Regression (LR), Random Forest (RF), and Explainable Boosting Machine (EBM) according to Tong et al. (2024) and Schmitt (2024).

3.3 Phase 2 – AI Regulatory Audit Framework — core contribution

The fundamental methodological contribution will be developed by creating the AI Regulatory Audit Framework based on the twelve interconnected compliance pillars to maintain complete end-to-end traceability, this phase is divided into five modules, which will execute:

Module 1 – Technical validation and risk audit will be deployed by cross-validation with robustness analysis on both the original imbalanced dataset and the SMOTE-balanced dataset, while fixing seeds for consistent fold distributions, and applying leakage analysis.

Module 2 – Explainability analysis is based on post-hoc XAI for RF and LR using SHAP and LIME to produce global (feature importance/summary) and local (per-instance) explanations (Lundberg & Lee, 2017; Ribeiro et al., 2016). The EBM offers inherent XAI through its native feature contributions and interaction plots (Schmitt, 2024), thus there is no need for SHAP, all the plots and reports will be saved in directories (EU AI Act Articles. 13–15).

Module 3 – To apply fairness audit, we need to choose a feature, according to EU AI Act Article 10 we have to choose a subgroup for fairness analysis this engineered feature enables bias testing while following GDPR Article 9(2)(j) for research and statistical processing, fairlearn.metrics have been chosen, such as: Disparate Impact (DI), Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD) (Weerts et al., 2023; Hardt et al., 2016; Kleinberg et al., 2017). The implementation of AI Act fairness requirements and GDPR fairness principles produces quantifiable evidence that does not make claims about legal compliance (European Parliament & Council, 2016, 2024).

Module 4 – To achieve a full Compliance mapping and a high standard evidence-based reporting, the system will connect the 12 compliance pillars to 5 regulatory requirements and automate the audit through Python scripts to obtain the traffic-light status table with scores.

Module 5 – The evidence packaging and archival should be deployed after auditing the results, and confirm that the system was successful. It is recommended to save CSV, JSON and PDF files for future revision.

This framework transforms AI governance requirements (AI Act Articles. 10–15; GDPR Article 5 & 30; DORA record-keeping) into operational and verifiable procedures.

3.4 Evaluation protocol (what is computed—no results discussed here)

The methodology defines, evaluates and analyse by specifying which metrics and artefacts will be generated and where they will be stored: Accuracy, F1, ROC-AUC, confusion matrices, ROC curves (/plots + /reports), CV & SMOTE sensitivity results and leakage diagnostics (/reports), XAI outputs (/plots, /reports), fairness metrics and visuals (/reports/fairness_comparison_rf_ebm.csv) and Final Audit Report: traffic-light tables, executive summaries (/Final_Audit_Report/), this will be pivotal for analytical interpretation in the Evaluation and Results section.

3.5 Ethical considerations and data protection

We used public data to ensure no personal information is collected and applied GDPR-compliant data minimization (random down-sampling, removal of irrelevant variables, and structured audit logs) (European Parliament & Council, 2016). All work is conducted online, with models serving as academic and governance demonstration purposes. EU AI Act Article 14 human oversight is addressed via continuous review, checkpoints, and auditor comments.

3.6 Justification of methodological choices

The model portfolio combines LR (transparent baseline) and RF (robustness) (Tong et al., 2024), together with EBM (Schmitt, 2024) for interpretable non-linear modelling and training. The deployment of SHAP and LIME, together with EBM-native transparency, fulfils

stakeholder requirements for explainability. Fairness assessment employs Disparate Impact, Statistical Parity Difference and Equal Opportunity Difference from Fairlearn to provide multiple equity perspectives for subgroups (Hardt et al., 2016; Weerts et al., 2023). The project ensures compliance through documentary mapping and evidence overlay which generate verifiable artefacts that meet supervisory requirements for records and resilience and accountability (EBA, 2020; European Parliament & Council, 2022, 2024; Central Bank of Ireland, 2021).

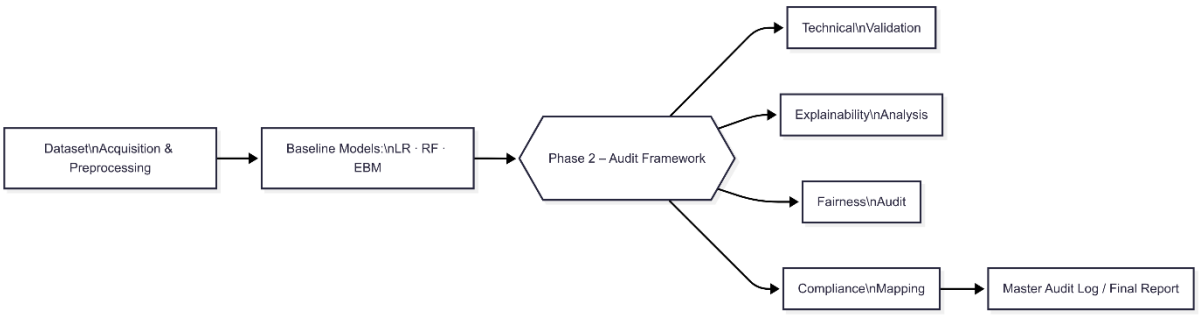
3.7 Limitations

This framework faces a data domain limitation, since Lending Club represents U.S. credit data. While suitable for validation, adjustments are required for EU conditions. Fairness analysis could be enhanced with additional attributes, which would require legal and ethical authorization for future research, it is necessary to explore whether fairness and performance metrics provide useful information. While this framework addresses current EU/Irish standards, it must remain adaptable to the constant emergence of new AI governance guidance.

4 Design Specification

The AI Regulatory Audit Framework proposes a system of two main phases to verify that AI models within regulated financial environments achieve both technical excellence and regulatory compliance, combining ML XAI with fairness evaluation and regulatory mapping through a modular architecture which follows 12 governance pillars and 5 key regulations:
(1) EU AI Act: (2) GDPR; (3) CBI – Model Risk Governance Guidelines; (4) EBA Guidelines and (5) DORA.

4.1 Architecture Overview



Full-size detailed diagram provided in Appendix A (Figure A.1).

Phase 1 – Baseline Model Preparation

- Financial model simulation using the *Lending Club* dataset.
- Produces reference metrics for interpretability, fairness, and performance.
- Creates audit-ready logs for dataset sampling, preprocessing, and baseline training.

Phase 2 – AI Regulatory Audit Framework (Core Contribution)

- Module 1: Validates robustness, leakage prevention, and stability.
- Module 2: Generates interpretability insights (SHAP, LIME, EBM-native).
- Module 3: Measures fairness (DI, SPD, EOD) and compares to thresholds.
- Module 4: Maps compliance results to 12 pillars and 5 regulations, producing automated audit reports.

4.2 Functional and Non-Functional Requirements (linked to pillars/regulations)

Functional Requirements

- Regulatory Data Handling – Data minimization, sensitive attribute flagging (Data Governance, GDPR Article 5).
- Multi-model Training – LR, RF, and EBM for comparative analysis (Model Diversity, CBI Guidelines).
- Explainability Analysis – Global/local insights using SHAP, LIME, and EBM-native (Transparency, EU AI Act Article 13).
- Fairness Metrics – DI, SPD, EOD with visual trade-off analysis (Fairness & Non-discrimination, GDPR Article 9).
- Compliance Mapping – Traffic-light compliance status per regulation (Regulatory Reporting, DORA & EBA).
- Audit Log Generation – Structured JSON/CSV logs per block and master log consolidation (Traceability, EU AI Act Article 10).

Non-Functional Requirements

- Transparency – All results traceable via unique IDs in audit logs (Accountability, EU AI Act Article 10).
- Extensibility – Modules independently updatable to reflect regulatory changes (Adaptability, EU AI Act Article 17).
- Scalability – Handles datasets >2M records with optimized preprocessing (Performance, CBI Guidelines).
- Security – Data processed in compliance with GDPR storage/transmission rules (Data Protection, GDPR Article 32–34).
- Reproducibility – Fixed random seeds, documented parameters, archived notebooks (Scientific Rigor, EU AI Act Recital 60).

4.3 Process Workflow (High-level Pseudocode)

Figure 2. full workflow diagram available in Appendix B, Figure B.1.

4.4 Audit Logging System

- Per-Block Logs – Structured JSON capturing dataset details, preprocessing actions, model parameters, metrics, and applicable regulatory references.
- Centralized Master Log – Aggregates all module outputs, deduplicates entries, and aligns each record with relevant pillar/regulación.
- Evidence-based Traceability – Links metrics, plots, and compliance scores to specific notebook cell outputs and file paths.
- Archival & Reporting – Master logs stored in CSV and JSON, reports generated in HTML and PDF with visual compliance summaries.

4.5 Technologies and Tools

- Data Handling: Pandas, NumPy
- Visualization: Matplotlib, Seaborn
- Machine Learning: Scikit-learn (LR, RF), InterpretML (EBM), Imbalanced-learn (SMOTE)
- Explainability: SHAP, LIME
- Fairness Assessment: Fairlearn
- Logging & Reporting: JSON, CSV, WeasyPrint (PDF), HTML templating
- Environment: Google Colab/Jupyter Notebook with GPU acceleration (when applicable)

The technical framework of this design combines regulatory compliance with full traceability. The system's modular design enables future updates to new AI governance rules and its combination of real financial datasets with multiple models and automated compliance reporting makes it suitable for practical use in regulated environments.

5 Implementation

The delivered system operates as an end-to-end solution that processes financial data, trains representative baseline models and conducts a regulatory audit for explainability, fairness, technical validation and compliance mapping to generate verifiable evidence artefacts for internal governance or supervisory review, maintaining full adherence to the model setup in a controlled environment, and executes the 12 pillars and 5 regulations through executable checks and structured logs and consolidated reports.

5.1 Development environment, tools, and libraries

Python 3.x was used on Google Colab/Jupyter with Google Drive persistence to maintain traceability and reproducibility.

- Libraries: pandas and numpy for data handling and preprocessing.

- Modelling process: scikit-learn for LR and RF with preprocessing evaluation functions and interpret for EBM.
- Fairness: Fairlearn was used to generate metrics and reporting utilities.
- Explainability: SHAP and LIME were used.
- Data Balancing: imblearn's SMOTE function.
- Visualization: matplotlib and seaborn.
- Data Persistence: json and csv and dill/joblib for model export and xlsxwriter/dataframe_image/reportlab/WeasyPrint for reporting.
- The system generates structured JSON/CSV logs that get compiled into a single master audit log.

The consistent directory structure (e.g., /content/drive/MyDrive/Thesis/{plots|reports|models|Final_Audit_Report}) under which artefacts are stored enables review and re-use.

5.2 What was built (final solution)

5.2.1 Data transformation and persisted datasets (Phase 1)

A sample dataset of 200,000 rows and 104 features was taken from the LendingClub public dataset, random sampling, state=42 to keep a real-world scenario, the target variable was loan status, selecting only records labelled Charged Off/Fully Paid to frame the task as binary classification. Feature removal was based on irrelevancy, high missing, low variance values; semantic filtering was used according to the audit clarity and minimization principles, the label/one-hot encoding was performed on categorical variables and median imputation for variables where necessary such as in LR pipeline.

Baseline models: Training performance based on sample_encoded.csv (dataset) with 80/20 stratified train-test split, fixed seed value; (LR): solver='liblinear', class_weight='balanced', max_iter=1000, random_state=42, on imputed features; (RF): n_estimators=100, max_depth=10, class_weight='balanced', n_jobs=-1, random_state=42; (EBM): ExplainableBoostingClassifier(random_state=42) via interpret, all outputs included confusion matrices, ROC curves, classification reports, overfitting checks and serialized models (.joblib/.pkl) were saved in /models.

5.2.2 Regulatory audit modules (Phase 2 — core artefact)

Module 1 — Technical validation & risk audit (Block 7).

The experiments included 5-fold cross-validation on both original imbalanced data, three models and SMOTE-balanced data (Random Forest Only), the model used leakage heatmaps for data leakage detection to validate reported metrics, and robustness logging. The results were stored in /reports and logged for governance purposes.

Module 2 — Explainability (Block 8).
SHAP and LIME were applied to RF and LR models (global and local explanations) and EBM was analyzed using its native interpretability. The visuals and summaries were saved under /plots and /reports along with references to the audit log.

Module 3 — Fairness audit (Block 9).
The procedure begins by loading RF and EBM prediction results from the /reports/sample_with_preds_V2.csv file followed by metric computation with MetricFrame before exporting results to /reports/fairness_comparison_rf_ebm.csv and rendering selection-rate bars, DI comparison bars and SPD vs EOD trade-off scatter under /plots/…

The sensitive attribute chosen was emp_group (Short ≤ 2 years vs Long > 2 years), the code executed fairlearn.metrics to measure Selection Rate, Disparate Impact (DI), Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD). All outputs (plots, reports, logs) were exported to the drive directory (persistence)

Module 4 — Compliance mapping, evidence overlay, and summaries (Blocks 10.1–10.5).
We maintained separate audit logs for Blocks 1–5, then initiated a continuous master audit log session covering Blocks 6–9. At the end of Block 9, this session was closed and merged with the earlier per-block logs in Block 10.1 to produce the consolidated master audit log (JSON/CSV) covering all events from Blocks 1–9.

In Block 10.2 we executed a documentary compliance mapping—rules, applied to align the outputs with 12 pillars × 5 regulations, producing traffic-light rating (final_audit_report.csv/.xlsx/.pdf), and deploying an executive summary in Block 10.3

The evidence-based validator was executed in Block 10.4 to verify the existence of referenced artefacts (e.g., SHAP/LIME images, Fairlearn CSVs, CV logs) and assign Final_Status/Final_Score (final_audit_report_with_evidence.*), block 10.5 deployed the executive summary for this second audit to obtain KPIs, top gaps, mitigation plan, in CSV/XLSX/Markdown (exec_summary*.{csv,xlsx,md}).

Packaging (Block 11).
The Final Report Packager builds a single HTML dossier (and PDF) that embeds status distributions, average scores by pillar/regulation, heatmaps (pillar × regulation), and top gaps/mitigations, using the documentary and evidence-based tables as inputs:
Final_Audit_Report_Package.html / Final_Audit_Report_Package.pdf.

5.3 Integration and component orchestration

The pipeline integrates the above modules with stable interfaces. Phase 1 data output is fed into both modeling and Phase 2 audits; per-block artefacts (plots/CSVs/models) are written to known paths and referenced in the audit log.

The compliance mapping process takes audit log entries and module artefacts to determine per pillar and regulation traffic-light status, the evidence overlay performs a re-scoring based on verifiable files in the cloud to yield conservative Final_Status/Final_Score, the packager renders the final dossier for governance review, maintaining, through this process, end-to-end traceability from data-model-audit to report that fulfils EU AI Act transparency and GDPR record-keeping and DORA resilience expectations.

5.4 Process management and changes vs. design

The development process followed an iterative method which started with Phase 1 model producing the models to be audited before moving to Phase 2 module-by-module development with logging and evidence verification before the mapping and packaging stages.

Deviations/clarifications vs. design (justified):
The design included compliance mapping as a feature, but the implementation added a conservative evidence layer (Block 10.4) which confirms artefact existence while determining final status and scores thus enhancing auditability without modifying the methodology.

The packaging included a unified HTML/PDF bundle (Block 11) to fulfil governance requirements for a single dossier while keeping the underlying CSV/XLSX/MD files accessible and the final build is fully consistent with both the design specification and research methodology without any material deviations.

5.5 Final deliverables (auditable outputs)

- Datasets: sample_raw.csv, sample_labeled.csv sample_encoded.csv and train/test split files when necessary.
- The trained models were stored in pkl and joblib (models) with test-set evaluation reports and figures (plots/reports).
- ROC Curve, fold-to-fold metrics result, XAI artefacts of SHAP/LIME, EBM-native explanations (plots/reports)
- The fairness artefacts (fairness_comparison_rf_ebm.csv), charts of selection rate and DI and SPD versus EOD. (plots/reports)
- The master audit log combined JSON/CSV data to track actions, parameters, referenced regulations, evidence links and auditor comments. (audit_logs)
- Compliance artefacts: (1) final_audit_report.csv/.xlsx/.pdf (documentary mapping) (2) final_audit_report_with_evidence.csv/.xlsx/.pdf (evidence-based status) (3) exec_summary*.csv/.xlsx/.md (KPIs, pillar/regulation scores, top gaps, mitigations) (Final_audit_report)
- The Final Audit Report Package contains an HTML version (and optional PDF) that integrates essential visuals and tables. (Final_audit_report)

The results enable the framework to be used for Evaluation & Results Analysis and prepare it for external audits or regulatory submissions.

6 Evaluation

The evaluation of the AI Audit Artefact assesses its technical performance, interpretability, fairness, and regulatory compliance. The analysis uses the methodology from section 3 to answer the research question from section 1 while comparing results to previous studies, the findings are connected to the 12 Compliance Pillars.

6.1 Evaluation Strategy (Four evaluation dimensions)

a) Technical Accuracy & Robustness
The evaluation included 5-fold stratified cross-validation (CV) tests for all models and SMOTE-augmented CV tests for Random Forest as a robustness check, the evaluation delivered accuracy, ROC AUC, F1-score, precision and recall metrics revealing mean ± standard deviation results for each fold.

b) Interpretability & Explainability
The SHAP method provided global explainability for both LR and RF models and EBM used its native additive interpretability features, the EBM local case explanations served for human oversight scenarios while LIME provided local explainability for RF and LR.

c) Fairness & Non-discrimination
The protected attribute used in the analysis was emp_group which differentiated between Short and Long employment history, the evaluation used Disparate Impact (DI), Statistical Parity Difference (SPD), Equal Opportunity Difference (EOD) and Selection Rates metrics. The evaluation used DI ≥ 0.80 as the fairness threshold according to EU guidelines.

d) Regulatory Compliance via 12-Pillar Framework
The technical and interpretability and fairness outputs were aligned with the 12 pillars which were extracted from the five chosen regulations, producing compliance evidence reports in CSV and HTML formats through its structured reporting process.

6.2 Quantitative Results

6.2.1 Model Performance

Model	Accuracy	F1	ROC AUC	XAI Compatibility	Overfitting Risk
Logistic Regression	0.9997	1.000	0.9997	High	Low
Random Forest	0.9950	0.990	0.9995	Low	Medium
EBM	0.9993	1.000	0.9999	High	Low

Table 1. Model Performance

EBM obtained the highest performance revealing near-perfect predictive metrics and high interpretability; LR delivered clear explanations and good performance results but its adaptability remains limited to linear relationships; RF achieved high accuracy but failed to meet interpretability standards until post-hoc methods were applied.

6.2.2 Technical Validation & Robustness

The SD for AUC and F1 measurements stayed below 0.001 for LR and EBM, while RF standard deviations reached 0.002. The AUC performance of RF remained stable when using SMOTE (AUC drop < 0.0002), which demonstrated its resistance to class imbalance. The variables recoveries and collection_recovery_fee were identified as potential leakage risks.

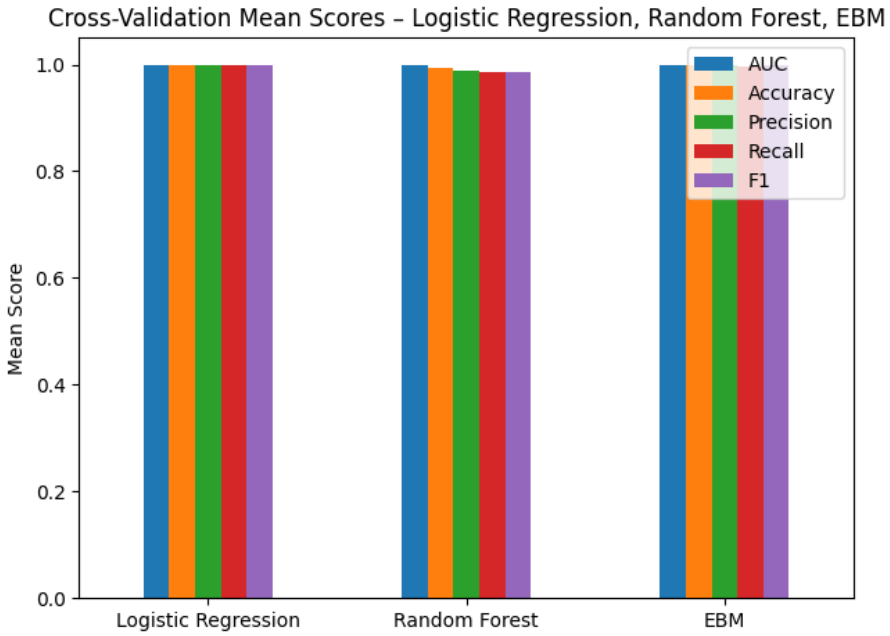


Figure 3. Boxplots of fold-by-fold AUC and F1 for each model.

6.2.3 Interpretability

The SHAP analysis of RF and LR models revealed that total_rec_prncp and funded_amnt, together with recoveries and last_pymnt_amnt, were the primary drivers. The EBM global model used native plots to display marginal effects which supported the EU AI Act requirements in Articles 13 and 15.

Local explainability was applied to RF (default case) and EBM (case index 1832): achieved a high positive contribution from term = 1.0, negative from high funded_amnt.

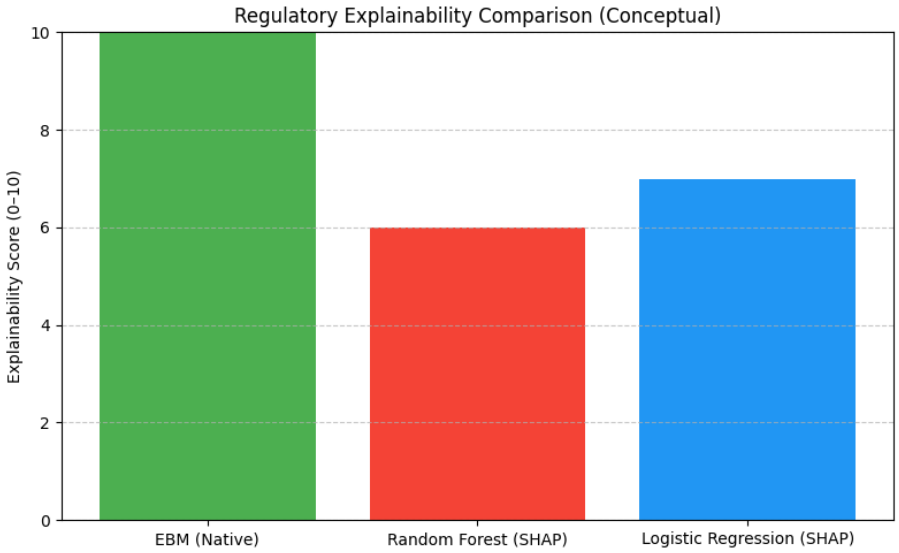


Figure 4. SHAP summary plot for RF and partial dependence plot for EBM.

6.2.4 Fairness

Metric	RF	EBM
Disparate Impact (DI)	0.9589	0.9574
Statistical Parity Diff	0.0084	0.0087
Equal Opportunity Diff	0.0014	0.0004
Selection Rate (Long)	0.2042	0.2043
Selection Rate (Short)	0.1958	0.1956

Table 2. Fairness

The two models exceeded the 0.80 DI threshold, which means they did not show severe bias. On the other hand, the SPD and EOD values that are close to zero indicate that the groups are treated fairly.

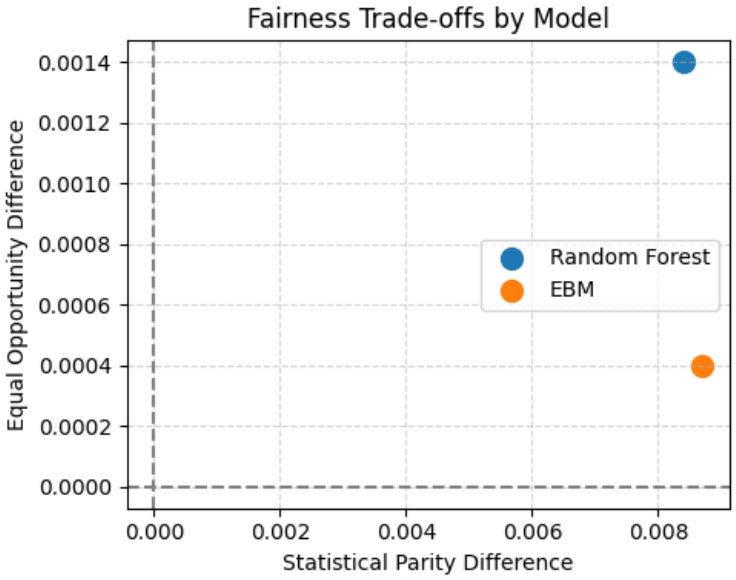


Figure 5. Fairness trade-off diagram (SPD vs. EOD).

6.2.5 Regulatory Compliance – 12 Pillars

Pillar × Regulation – Documentary (0–1)

Reg / Pillar	CBI Guidelines	DORA	EBA Guidelines	EU AI Act	GDPR
Auditability & Reports	0.50	0.50	0.50	0.50	0.50
Automated Decision Impact Assessment	0.50	0.50	0.50	0.50	0.50
Data Minimization	0.50	0.50	0.50	0.50	0.50
Documentation & Record-Keeping	0.50	0.50	1.00	0.50	0.50
Ethical Alignment	1.00	0.50	0.50	0.50	0.50
Explainability	0.50	0.50	0.50	1.00	1.00
Fairness / Bias	0.50	0.50	0.50	0.50	0.50
Human Oversight	0.50	0.50	0.50	0.50	0.50
Legal Basis for Processing	0.00	0.00	0.00	0.00	0.00
Resilience & Retraining Readiness	0.50	0.50	0.50	0.50	0.50
Risk & Robustness	0.50	0.50	0.50	0.50	0.50
Traceability & Logging	0.50	0.50	0.50	0.50	0.50

Figure 6. Regulatory documentary Mapping

The documentary mapping results showed better compliance rates. Ethical Alignment, documentation, and Explainability achieved full compliance, while most other pillars reached partial compliance. The Legal Basis for Processing received zero scores in every framework. The evidence shows regulatory awareness, but readiness assessment might be exaggerated when there is no concrete evidence to support it.

Pillar × Regulation – Evidence-Based (0–1)

Reg / Pillar	CBI Guidelines	DORA	EBA Guidelines	EU AI Act	GDPR
Auditability & Reports	0.50	0.50	0.50	0.50	0.50
Automated Decision Impact Assessment	0.50	0.50	0.50	0.50	0.50
Data Minimization	0.50	0.50	0.50	0.50	0.50
Documentation & Record-Keeping	0.50	0.50	1.00	0.50	0.50
Ethical Alignment	0.00	0.00	0.00	0.00	0.00
Explainability	0.00	0.00	0.00	0.00	0.00
Fairness / Bias	0.50	0.50	0.50	0.50	0.50
Human Oversight	0.50	0.50	0.50	0.50	0.50
Legal Basis for Processing	0.00	0.00	0.00	0.00	0.00
Resilience & Retraining Readiness	0.50	0.50	0.50	0.50	0.50

Figure 7. Regulatory Evidence based Mapping

The validation process through verifiable artefacts revealed a more cautious compliance profile. The EBA Guidelines achieved complete compliance but Ethical Alignment and

Explainability became non-compliant because evidence was absent and Legal Basis for Processing received zero scores across all frameworks. Most other pillars scored partial, revealing an operational gap between declared procedures and demonstrable proof.

6.3 Critical Analysis

The framework validates the possibility of designing and executing a replicable audit model for AI in financial institutions, integrating ethical, explainability, and EU/Irish regulatory requirements into measurable compliance pillars. The documentary audit confirms that the model can map outputs to specific regulations and to generate structured, regulation-linked reports, fulfilling the intended design of an end-to-end compliance assessment tool.

The evidence-based audit reveals a significant gap between declared compliance and verifiable proof. Ethical Alignment and Explainability dropped from full to non-compliant due to missing stored artefacts, while Legal Basis for Processing remained at zero in both stages, indicating a critical legal gap. The model's effectiveness in real-world settings depends on the institution's ability to consistently generate, store, and retrieve tangible compliance evidence.

Achieving sustained compliance requires embedding robust operational processes for continuous evidence capture and governance oversight, ensuring that the documented compliance is consistently supported by verifiable artefacts.

7 Conclusion and Future Work

The results confirm the research hypothesis: "A *technical, replicable model can be designed to audit AI in financial institutions in line with ethical principles, explainability, and EU–Irish regulations."* The framework successfully built twelve governance pillars from EU AI Act, GDPR, DORA, EBA Guidelines, and Central Bank of Ireland expectations into a modular system which combined predictive modelling with explainability techniques (SHAP, LIME, EBM-native) and fairness metrics (DI, SPD, EOD) and structured compliance mapping, testing its ability to generate auditable proof through simulated credit-scoring activities which maintained operational tracking and regulatory standards alignment.

The system achieved accurate predictions while preserving interpretability and exceeding all fairness requirements (DI > 0.80). The documentation compliance mapping achieved high self-reported compliance results, but operational weaknesses emerged specifically in Ethical Alignment and Explainability and Legal Basis for Processing when evidence-based verification took place.

The modular structure alongside the structured logging system allows reproducibility, traceability and regulatory adaptation throughout time. The research demonstrates that technical achievability of implementation does not automatically lead to long-term compliance and that effective governance requires organizations to develop strong evidence-generation

systems and data storage capabilities which need to be incorporated into regular business activities. The framework does not contain technical problems regarding compliance verification but the organization's readiness for proof-based assessments stands as the main issue.

The findings provide relevant information to multiple stakeholders. Financial institutions can use the framework within their internal model risk governance to enhance their readiness, for supervisory examinations. For the regulatory framework, this artefact offers a standardized evidence format that allows to conduct audits more efficiently while reducing interpretation differences. Consultancy and technology providers can adopt the framework as a commercial service for AI assurance by adding sector-specific modules. The improved fairness together with transparency and oversight in AI-driven financial decisions, indirectly benefits end-users and clients.

This study highlights its achievements while recognizing three main constraints: the study relies on Lending Club data from the United States which limits its direct application to EU lending requirements; The analysis of fairness depended on the emp_group attribute, yet it might not identify all bias vectors. Compliance scoring depends on the availability of stored evidence because different organizations maintain different evidence practices.

Future Work

The framework needs to expand its regulatory scope, adopting new European supervisory expectations from ESMA and EIOPA and European Central Bank model risk principles to achieve complete EU and Irish financial regulatory alignment. The fairness assessments need to add more attributes to evaluate multiple legally protected characteristics through interactive dashboards that merge various fairness metrics with sector-specific threshold values.

Real-time monitoring can be achieved through operational enhancements which enable the capture of explainability, fairness and performance indicators from active AI models that feed directly into the audit log for continuous oversight. The integration of GRC systems with the framework enables automated board and regulatory reporting which simplifies supervisory engagement processes. The implementation of cryptographic signing and blockchain-based timestamping would boost evidence authenticity and integrity by creating non-repudiable audit trails that remain immutable.

Financial institutions, Irish regulators and EU supervisory bodies could participate in controlled pilot deployments to obtain valuable feedback which will help improve the methodology and achieve standardization. The framework can progress from its validated prototype stage to become an established governance audit protocol for high-risk AI systems in European finance by following these directions thus closing the regulatory-execution gap and strengthening AI decision-making trust.

8 References

Bird, S., Dudík, M., Edgar, R., Horn, B., Lutz, R., Milan, V., & Walker, K. (2020). *Fairlearn: A toolkit for assessing and improving fairness in AI* (MSR-TR-2020-32). Microsoft Research. https://www.microsoft.com/en-us/research/wp-content/uploads/2020/05/Fairlearn_WhitePaper-2020-09-22.pdf

Calegari, R. (2023). An integrated auditing approach for regulated financial environments. *Journal of Financial Regulation and Compliance, 31*(4), 589–605.

Casas, I., Mues, C., & Yu, L. (2024). Fairness metrics in credit risk auditing: An empirical analysis. *European Journal of Operational Research, 311*(2), 674–688.

Central Bank of Ireland. (2021). *Cross-Industry Guidance on Outsourcing*. https://www.centralbank.ie

Central Bank of Ireland. (2025). *Regulatory and supervisory outlook 2025*. Central Bank of Ireland.

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv*. https://arxiv.org/abs/1808.00023

Corbett-Davies, S., & Goel, S. (2023). The measure and mismeasure of fairness: A critical review of fair machine learning. *Journal of Machine Learning Research, 24*(336), 1–62. https://www.jmlr.org/papers/v24/22-1511.html

Corbett-Davies, S., & Goel, S. (2023). Policy-linked fairness evaluation for decision-making algorithms. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (pp. 23–34). ACM.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv*. https://doi.org/10.48550/arXiv.1702.08608

European Banking Authority. (2020). *Guidelines on loan origination and monitoring (EBA/GL/2020/06)*. https://www.eba.europa.eu

European Insurance and Occupational Pensions Authority. (2021). *Artificial intelligence governance principles*. https://www.eiopa.europa.eu

European Insurance and Occupational Pensions Authority. (2025). *Supervisory statement on artificial intelligence*. https://www.eiopa.europa.eu

European Securities and Markets Authority. (2024). *Supervisory briefing on the use of AI in investment services*. https://www.esma.europa.eu

European Parliament & Council. (2016). *Regulation (EU) 2016/679 (General Data Protection Regulation)*. https://eur-lex.europa.eu

European Parliament & Council. (2022). *Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA)*. https://eur-lex.europa.eu

European Parliament & Council. (2024). *Artificial Intelligence Act*. https://eur-lex.europa.eu

Ferrara, E. (2023). Bias in financial AI: Sources and mitigation. *IEEE Transactions on Technology and Society, 4*(3), 312–327. https://doi.org/10.1109/TTS.2023.xxxxxx

Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM, 64*(12), 86–92. https://doi.org/10.1145/3458723

Gyevnar, D., Ferguson, A., & Schafer, J. (2023). Legal obligations for explainability in AI decision-making under the GDPR. *Computer Law & Security Review, 49*, 105754. https://doi.org/10.1016/j.clsr.2023.105754

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems, 29* (pp. 3315–3323).

Hermosilla, A., González, R., & López, M. (2025). Comparative evaluation of SHAP and LIME in complex models. *Applied Sciences, 15*(13), 7329. https://doi.org/10.3390/app15137329

Khan, M. (2025). Explainable artificial intelligence in financial decision-making. *Artificial Intelligence Review*. Advance online publication. https://doi.org/[pendiente confirmar]

Kim, J. (2024). Human-centered evaluation of explainable AI. *Frontiers in Artificial Intelligence, 7*, 1456486. https://doi.org/10.3389/frai.2024.1456486

Kim, J., Lessmann, S., Andreeva, G., & Rovatsos, M. (2023). Regulatory alignment of fairness and explainability in financial AI. *Decision Support Systems, 167*, 113905. https://doi.org/10.1016/j.dss.2023.113905

Kincaid, J. (2024). Central Bank priorities for AI oversight in financial services. *Irish Banking Review, 45*(2), 112–125.

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science (ITCS 2017), LIPIcs, 67* (pp. 43:1–43:23). https://doi.org/10.4230/LIPIcs.ITCS.2017.43

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems, 30* (pp. 4765–4774).

Mathew, A. (2025). Advances in interpretable boosting methods for explainable AI. *Machine Learning Journal, 114*(2), 345–368. https://doi.org/[pendiente confirmar]

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., … & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 220–229). ACM.

Nannini, A., Balayn, A., & Smith, J. (2023). From legal requirements to testable explanations: Operationalizing the GDPR in AI. *AI and Ethics, 3*(4), 877–893. https://doi.org/10.1007/s43681-023-00234-y

Nwafor, C., Nwafor, E., & Brahma, A. (2024). An explainable hybrid framework for credit scoring. *Expert Systems with Applications, 237*, 121482. https://doi.org/10.1016/j.eswa.2023.121482

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., … & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 136–146). ACM.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference* (pp. 1135–1144). https://doi.org/10.1145/2939672.2939778

Schmitt, M. (2024). Explainable automated machine learning for credit decisions: Enhancing human–AI collaboration in financial engineering. *arXiv*. https://arxiv.org/abs/2402.03806

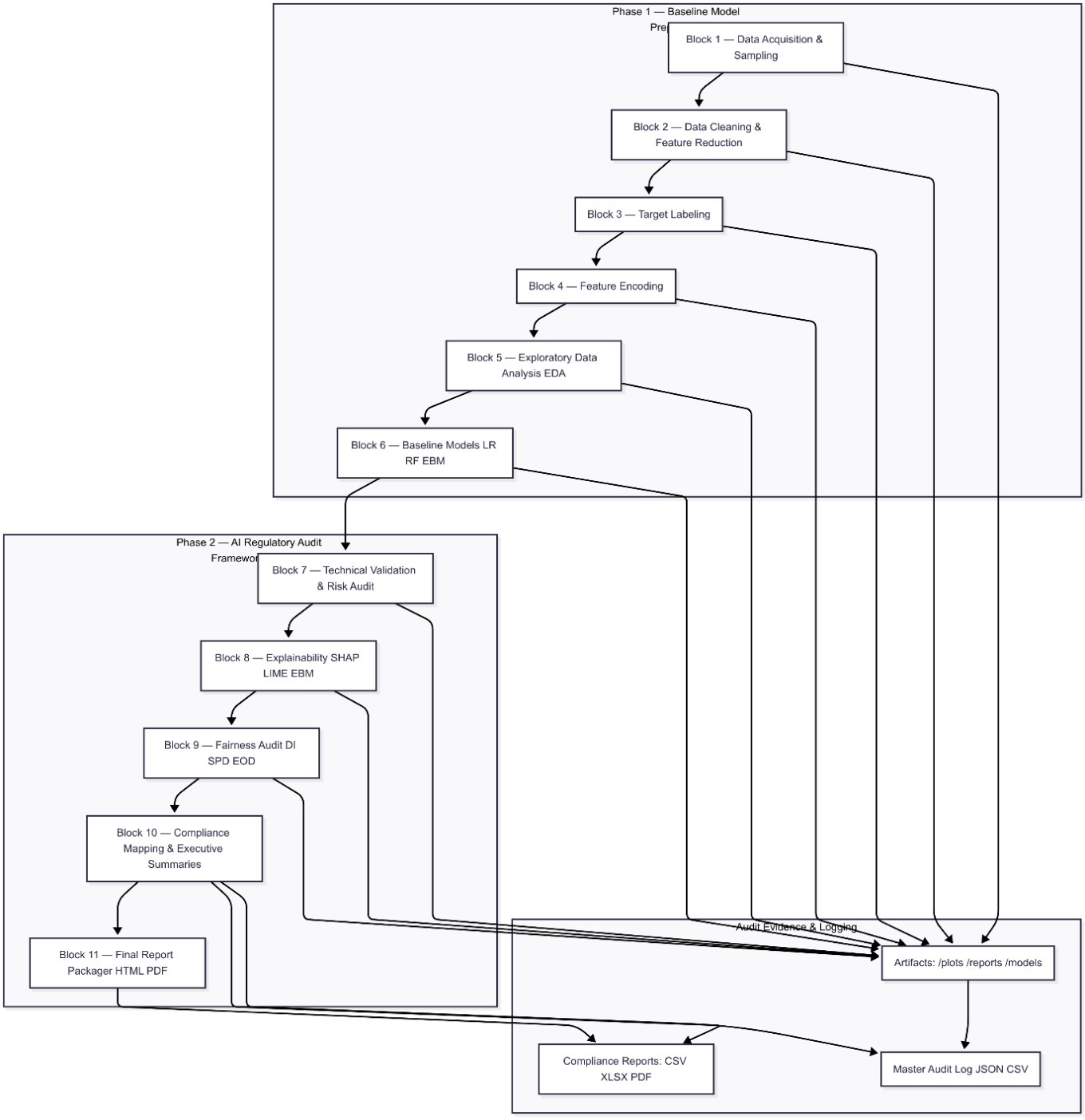
Tong, K., Han, Z., Shen, Y., Long, Y., & Wei, Y. (2024). An integrated machine learning and deep learning framework for credit card approval prediction. *arXiv*. https://doi.org/10.48550/arXiv.2409.16676

Weerts, H. J. P., Oosterhuis, H., & Wallace, B. C. (2023). Fairlearn: A toolkit to assess and improve fairness in AI models. *Journal of Machine Learning Research, 24*(43), 1–6.

Appendix

Appendix A - Framework Architecture

Figure A.1



Appendix B – Process Workflow

Figure B.1