GOVERNING CREDIT WITH DIGITAL TWINS: EXPLAINABLE AI, CREDIT OFFICER DIGITAL TWINS, AND THE EU AI ACT

Giammarco TOSI
"Aurel Vlaicu" University of Arad, Romania
E-mail: gmtosi@gmail.com
ORCID: 0009-0008-6864-0721

Abstract: *The deployment of artificial intelligence in credit decisioning has substantially improved predictive accuracy, yet the opacity of advanced machine learning models raises critical concerns about transparency, fairness, and regulatory compliance. The paper examines the intersection of Explainable AI (XAI), digital twin technology, and the European Union AI Act within the domain of credit origination and underwriting. A conceptual framework is proposed in which a digital twin of the credit deliberation process is complemented by a behavioural Credit Officer Digital Twin, together forming an integrated environment for real-time simulation, multi-level explainability, bias and efficiency monitoring, and regulatory audit. Drawing on recent literature, the paper synthesizes findings on SHAP, LIME, counterfactual explanations, fairness-constrained optimization, and financial and organizational digital twins to argue that twin-enabled XAI architectures can reconcile the tension between model performance, human judgement, and regulatory transparency. The Credit Officer Digital Twin is modelled as a data-driven replica of the human credit officer's decision policy, augmented with analytics that surface individual biases, efficiency patterns, and inconsistencies, thereby supporting meaningful human oversight rather than replacing it. The overall framework operationalizes key requirements of the AI Act for high-risk AI systems, including data governance, documentation, human oversight, and the right to explanation under both the AI Act and the GDPR. Implications for banking practice, supervisory policy, and future research on responsible, human-in-the-loop credit decisioning are discussed.*

Keywords: *Explainable Artificial Intelligence, Credit Scoring, Credit Officer Digital Twin, EU AI Act, Algorithmic Fairness, Responsible AI.*

Classification JEL: *E51, O33, C53.*

UDC: *[004.8:336.77]+[341.1(EU):004.8]* ***DOI:*** *https://doi.org/10.53486/ser2026.46*

1. Introduction

Credit decisioning, the process by which financial institutions evaluate, approve, or reject loan applications, has undergone a fundamental transformation in recent years. Traditional statistical methods such as logistic regression and linear discriminant analysis, once the backbone of credit scoring, are increasingly supplemented or replaced by machine learning (ML) algorithms including gradient boosting machines, random forests, and deep neural networks. These models achieve superior predictive performance, reducing default rates and expanding the capacity to assess borrowers with limited credit histories. However, their complexity introduces a fundamental problem: opacity. The very architectures that drive predictive gains - ensemble methods, non-linear transformations, high-dimensional feature interactions - render the decision-making logic inaccessible to human stakeholders, from the borrower denied a mortgage to the compliance officer tasked with audit.

This opacity is not merely an inconvenience. It generates legal, ethical, and operational risks. Borrowers who receive an adverse credit decision are, in many jurisdictions, entitled to an explanation of the factors influencing that outcome. Regulators demand documentation and auditability of models deployed in high-stakes financial decisions. Algorithmic bias - whether inherited from historical data, amplified by feature engineering, or introduced through proxy variables - can produce discriminatory outcomes against protected groups without detection in opaque systems. The European Union's AI Act (Regulation (EU)

2024/1689), which classifies AI systems used for creditworthiness assessment as "high-risk," imposes stringent requirements for transparency, data governance, human oversight, and ongoing monitoring.

Simultaneously, digital twin technology - the creation of dynamic virtual replicas of physical systems or processes, continuously synchronized with real-time data - has emerged as a powerful paradigm in manufacturing, healthcare, and, increasingly, financial services. Financial digital twins enable institutions to simulate market scenarios, stress-test portfolios, and model operational disruptions in a risk-free environment. Yet the application of digital twin concepts specifically to credit deliberation processes, and to the behaviour of human decision-makers within those processes, remains largely unexplored in the academic literature.

This paper proposes a novel conceptual framework, the Extended Digital Twins for the Credit Deliberation Process (EDT-CDP) in which two types of digital twins are combined: a digital twin of the credit deliberation process (DT-CDP) and a Credit Officer Digital Twin. The DT-CDP models the end-to-end workflow of credit origination and underwriting, integrating Explainable AI methods, algorithmic fairness monitoring, and AI Act compliance mechanisms into a unified simulation and governance architecture. The Credit Officer Digital Twin, by contrast, is conceived as a learning and coaching environment for the human underwriter: a data-driven replica of the officer's decision policy that continuously learns from their overrides and judgements, makes their heuristics and blind spots explicit, and feeds back tailored insights on bias, consistency, and efficiency. Rather than replacing the human, the twin is designed to help the officer improve over time, by exposing how their decisions compare to model recommendations, institutional policies, and fairness targets, and by providing a safe sandbox in which to experiment with alternative choices without affecting real customers.

The paper addresses different topics: (1) a systematic review of recent literature on XAI techniques applied to credit decisioning; (2) an analysis of the regulatory requirements under the AI Act and GDPR relevant to credit scoring; (3) a conceptual architecture for a DT-CDP that operationalizes explainability, fairness, and compliance by design; and (4) the introduction of the Credit Officer Digital Twin as a coaching-oriented construct for analysing and augmenting human decision-making in credit underwriting.

2. Literature Review

The growing use of AI in credit scoring is marked by a persistent tension between predictive accuracy and interpretability. Advanced models such as XGBoost, LightGBM, and deep neural networks can outperform traditional approaches like logistic regression by capturing complex, non-linear feature interactions, but they often do so at the cost of transparency. As Barredo Arrieta et al. (2020) note, explainability concerns the ability to render the internal logic of AI systems understandable to human users, a distinction that is particularly important in credit decisioning, where borrowers, regulators, and institutions must understand not only the outcome but also the reasons behind it.

Recent studies suggest that this trade-off is no longer absolute. Ogbuefi et al. (2025) show that post-hoc XAI techniques can be integrated into high-performing credit models without materially undermining predictive performance, while comparative research highlights the complementary strengths of SHAP and LIME. SHAP, introduced by Lundberg and Lee

(2017), offers theoretically grounded and relatively stable explanations suitable for audit and risk management, whereas LIME, proposed by Ribeiro, Singh, and Guestrin (2016), provides simpler local explanations that may be easier to communicate to customers and loan officers, albeit with greater instability. Pathi and Pothineni (2025) therefore argue for hybrid approaches capable of serving different stakeholder needs within the same credit pipeline. Beyond feature attribution, counterfactual explanations further enrich the XAI toolkit by showing applicants the minimal changes that could have led to a different outcome. Wachter, Mittelstadt, and Russell (2018) frame these explanations as particularly relevant for transparency, contestability, and recourse, and Takahashi, Shimizu, and Tanaka (2024) refine this logic through causal methods that avoid unrealistic recommendations. In the credit domain, Prade (2022) places counterfactual and recourse generation within a broader explanatory pipeline that also includes bias auditing and continuous monitoring.

Yet the literature also makes clear that explainability remains far from straightforward. Post-hoc methods can suffer from instability, infidelity, multiplicity, and stakeholder misalignment: LIME may generate different explanations for nearly identical cases (Pathi and Pothineni, 2025; Prade, 2022), surrogate explanations may diverge from the true decision logic (Barredo Arrieta et al., 2020; Sowmiya et al., 2024), multiple narratives may be constructed around the same prediction, and technically sound explanations may still fail to satisfy borrowers or compliance officers. The issue, then, is not simply to "add" explanation to AI-based credit scoring, but to embed explanation within a broader governance structure capable of supporting transparency, consistency, and practical accountability.

This need becomes even more pressing under the current European regulatory framework. The EU AI Act (Regulation (EU) 2024/1689) explicitly classifies AI systems used to assess the creditworthiness of natural persons as high-risk, thereby imposing obligations related to risk management, data governance, documentation, logging, transparency, human oversight, and robustness. Hacker and Eber (2025) emphasize that the scope of the Act is broad enough to include not only sophisticated machine learning models but also more conventional decisioning systems, extending regulatory exposure across much of the lending sector. In parallel, the GDPR and the AI Act together strengthen the right to explanation: under the GDPR, individuals are entitled to meaningful information about the logic involved in automated decisions, while Article 86 of the AI Act requires clear and meaningful explanations of the AI system's role and the main elements of the decision. Juliussen (2025) and Metikos and Ausloos (2025) show, however, that translating these legal principles into operational practice remains challenging, especially when explanations are unstable, overly technical, or insufficiently tailored to different audiences. Similar concerns arise in relation to algorithmic fairness.

Article 10 of the AI Act, together with broader non-discrimination principles in EU law, requires that credit scoring systems do not produce systematically biased outcomes. Nagaraj (2025) demonstrates that fairness constraints can be incorporated directly into model optimization, although this also reveals the well-known difficulty that fairness criteria may conflict with one another. Sowmiya et al. (2024), Barredo Arrieta et al. (2020), Pathi and Pothineni (2025), and Prade (2022) further show that biased data, problematic feature engineering, unstable explanations, and stakeholder misalignment can together obscure rather than reduce discrimination. It is in response to this combination of technical opacity, legal obligation, and governance complexity that digital twin technology becomes particularly relevant.

Defined by Grieves (2014) and expanded by Grieves and Vickers (2017) as a dynamic virtual representation continuously synchronized with real-world data, digital twins have already found applications in banking for stress testing, fraud detection, credit strategy simulation, and infrastructure resilience. Pattabhi (2025) conceptualizes these as Financial Digital Twins, while Chatterjee et al. (2024), FICO's Business Outcome Simulator, Lloyds Banking Group, and BCG's Tech in Banking 2025 report illustrate their growing practical relevance. Nevertheless, the existing literature remains focused on portfolio risk, operational systems, and fraud, rather than on the credit deliberation process as a governed socio-technical workflow. No published study, to the knowledge reflected in the present review, has proposed a digital twin specifically designed to replicate and govern the full sequence from application intake to scoring, human review, final decision, explanation, and audit. It is precisely this gap that the present paper addresses.

3. Methodology

The proposed Extednded Digital Twins of the Credit Decision Process (EDT-CDP) is a layered architecture consisting of six interconnected modules:

1. Process Replication Layer: A dynamic model of the entire credit deliberation workflow - data collection, scoring, policy rule application, human deliberation, decision output, and client communication. This layer mirrors the real-world process in real time, capturing decision flows, timing, exceptions, and overrides.
2. XAI Engine: An integrated module that generates multi-stakeholder explanations for each credit decision. The engine combines:
   - SHAP-based global and local explanations for audit and risk management;
   - LIME-based local explanations for customer-facing communication;
   - Counterfactual explanations for applicant recourse, describing the minimal changes that would alter the decision.
3. Fairness Monitoring Module: A continuous bias auditing system that evaluates model outputs across protected attributes (gender, ethnicity, age, geography) using multiple fairness metrics (demographic parity, equal opportunity, equalized odds). Alerts are triggered when disparities exceed predefined thresholds, and the module tracks fairness metrics over time to detect drift.
4. Scenario Simulation Engine: A sandbox environment where credit policy changes, model updates, macroeconomic shocks, or regulatory changes can be simulated before deployment. Users can test the impact of introducing ESG criteria into credit policy, adjusting risk thresholds, or deploying a new ML model, observing projected effects on approval rates, default rates, profitability, and fairness metrics.
5. Compliance and Audit Layer: An automated module that generates regulatory documentation aligned with AI Act requirements: technical documentation (Article 11), logging and record-keeping (Article 12), transparency notices (Article 13), and human oversight protocols (Article 14). This layer maintains a tamper-proof audit trail of every decision, explanation, and intervention.

6. Credit Officer Digital Twin: A complementary module that models the human credit officer as a data-driven, dynamic behavioural twin, continuously updated through observed decisions and contextual information. It captures decision thresholds, override patterns, and feature sensitivities, while producing bias and efficiency analytics and delivering feedback that strengthens consistency, learning, and meaningful human oversight.

The EDT-CDP operationalizes explainability at three levels, corresponding to different stakeholders and AI Act requirements:

- Level 1 – Applicant-facing explanations: When a credit application is rejected (or approved with conditions), the DT-CDP automatically generates a plain-language explanation drawing on LIME-based local attributions and counterfactual statements. Example: "Your mortgage application was not approved. The main factors were: (1) your current debt-to-income ratio of 48%, which exceeds our threshold of 40%; (2) your employment tenure of 8 months, which is below the 12-month minimum for this product. If your debt-to-income ratio were reduced to 40% or below, the decision would change to approval." This fulfills the AI Act Article 86 requirement for "clear and meaningful explanations".
- Level 2 – Underwriter and risk management explanations: The XAI engine provides SHAP-based feature attributions at both individual and portfolio levels, displayed through an explainability dashboard. Underwriters can inspect the contribution of each variable to a specific decision, compare it against the portfolio distribution, and identify cases requiring manual review. This supports the AI Act's human oversight requirement (Article 14).
- Level 3 – Regulatory and audit explanations: The compliance layer generates structured reports documenting model logic, training data characteristics, performance metrics, fairness audit results, and a complete decision log with associated explanations. These reports are designed to meet the technical documentation requirements of the AI Act (Articles 11–12) and can be produced on demand for supervisory inspections.

While the DT-CDP models the end-to-end credit deliberation process, a complementary construct is the credit officer digital twin: a data-driven, dynamic virtual replica of a human credit officer's decision-making behaviour, continuously updated from observed decisions and contextual information. Rather than merely approximating model outputs, the twin learns the officer's implicit decision policy, including typical approval and rejection thresholds, patterns of manual overrides of model recommendations, and sensitivity to specific features such as employment stability, collateral quality, or ESG-related flags.

Building on emerging work on digital twins for human capital and organizational decision-making, the credit officer digital twin serves not only as a mirror but as an augmented decision support entity. On top of the behavioural profile, the twin incorporates analytics modules that estimate (i) bias indicators (for example, disparities in approval rates across demographic or geographic segments after controlling for risk-relevant variables), and (ii) efficiency metrics (such as time spent per case, intra-officer consistency on similar applications, and divergence from institution-wide policies). In operation, the twin can provide the officer with real-time feedback - highlighting when a current decision deviates from their usual pattern, from the model's recommendation, or from fairness and consistency targets - and in simulation mode it

allows the institution to explore how changes in policies, training, or incentives would reshape individual decision behaviour. This conception aligns with the AI Act's emphasis on meaningful human oversight: the officer remains in the loop, but is systematically augmented by a digital environment that surfaces their own blind spots, biases, and efficiency gaps, thereby improving both individual judgement and institutional governance.

4. Results and Discussion

The EDT-CDP framework contributes to the literature by reframing explainability in credit scoring from a predominantly model-centric problem to a broader question of socio-technical governance. Much of the existing XAI literature has concentrated on the interpretability of individual predictions, while the digital twin literature has focused on simulation, synchronization, and process optimization at the system level. The contribution of the present framework lies precisely in integrating these two perspectives. Rather than treating explanation as an add-on to an otherwise opaque model, the EDT-CDP embeds explainability within the architecture of the credit deliberation process itself, where model outputs, human intervention, fairness controls, and audit requirements are connected within a continuously updated digital environment. This shift is theoretically important because many of the problems identified in the credit-XAI literature - instability of post-hoc explanations, weak alignment with stakeholder needs, and limited integration with organizational routines - cannot be fully resolved through model-level interpretability alone. They require a governance structure capable of situating explanations within the wider institutional, procedural, and regulatory context in which credit decisions are made.

A second contribution lies in the framework's explicit operationalization of the AI Act's requirements for high-risk credit systems. Existing approaches often conceptualize compliance as an external layer added after model development, typically through documentation, validation procedures, or isolated explanation tools. By contrast, the EDT-CDP translates transparency, human oversight, fairness monitoring, logging, and documentation into native functions of the decision infrastructure. In doing so, it advances a compliance-by-design logic that is especially relevant for banking, where legal exposure increasingly depends not only on predictive performance but also on the demonstrability of traceability, contestability, and institutional control. This gives the framework significance beyond the technical domain of explainable AI: it also speaks to the growing policy debate on how high-risk AI systems should be governed in practice. The framework suggests that regulatory alignment is unlikely to be achieved through fragmented controls or ex post justification alone; instead, it requires integrated architectures in which explanation, monitoring, and accountability are built into the operational fabric of decision-making. In this sense, the EDT-CDP offers a conceptual bridge between the legal language of the AI Act and the organizational realities of model deployment in financial institutions.

A third and more original contribution is the introduction of a multi-level and multi-stakeholder logic of explanation within a digital twin environment. Prior research has repeatedly shown that explanations are not universally meaningful: what is useful for a model validator may be unintelligible to a borrower, while a borrower-oriented explanation may be too superficial for supervisory or audit purposes. The EDT-CDP addresses this limitation by conceptualizing explanation as a differentiated governance funcion rather than as a single output. Explanations become context-sensitive artifacts tailored to the

informational role of the recipient: recourse-oriented and plain-language explanations for applicants, model-based and portfolio-level insights for underwriters and risk managers, and structured traceability for auditors and regulators. This layered approach is important because it aligns explainability with institutional plurality rather than technical uniformity. It also reinforces the idea that the legitimacy of AI-supported credit decisioning depends not merely on whether an explanation can be generated, but on whether the explanation is appropriate, stable, and actionable for the audience that receives it.

From a managerial standpoint, the EDT-CDP offers financial institutions a credible response to a set of pressures that are often perceived as mutually constraining. Banks increasingly rely on complex machine learning models to enhance predictive accuracy and expand credit assessment capabilities, yet these same models intensify demands for transparency, fairness, and human accountability. The framework indicates that these objectives need not be treated as irreconcilable trade-offs if they are addressed through an integrated governance architecture. By combining XAI, fairness monitoring, simulation capabilities, and compliance functions in a single environment, the EDT-CDP allows institutions to move from reactive compliance - where explanations are generated only after a decision is challenged - to proactive governance, in which decision processes are explainable, reviewable, and auditable by construction. This proactive orientation is particularly relevant in credit markets, where reputational risk, litigation risk, supervisory scrutiny, and public concerns about discrimination increasingly converge. In such settings, the ability to demonstrate not only model validity but also procedural fairness and oversight quality becomes a strategic capability rather than a mere compliance obligation.

The framework also has important implications for organizational learning and resilience. One of the main limitations of conventional XAI tools is that they are largely retrospective: they help explain what happened in a given case, but they do not necessarily help institutions understand how future decisions may evolve under changing policies, data conditions, or macroeconomic shocks. The digital twin logic extends the role of explainability from retrospective justification to anticipatory governance. Through simulation, the institution can assess ex ante how changes in thresholds, policy rules, model updates, ESG criteria, or external market conditions may affect approval rates, fairness profiles, operational bottlenecks, and the stability of explanations themselves. This capability is significant because it repositions governance from a defensive function to a strategic one. Rather than merely documenting or correcting problematic outcomes after deployment, banks can test interventions before they affect real customers, thereby reducing operational and ethical risk while strengthening adaptive capacity.

More broadly, the EDT-CDP contributes to the discussion on human oversight by challenging minimalist interpretations of the “human in the loop” principle. In many real-world settings, human oversight is reduced to a formal possibility of override, without ensuring that the human actor is actually equipped to understand the model’s logic, detect problematic outputs, or recognize their own biases. The framework suggests that meaningful oversight requires a richer informational infrastructure, one that makes model outputs interpretable, deviations visible, trade-offs explicit, and interventions traceable. In this respect, the EDT-CDP redefines oversight not as a procedural formality but as an organizational capability supported by technical design.

This is particularly relevant when considered alongside the complementary role of the Credit Officer Digital Twin, which further extends the governance logic from process replication

to the augmentation of human judgment itself. Taken together, these elements point toward a broader reconfiguration of AI governance in credit decisioning: one in which explainability, fairness, simulation, and human oversight are not separate compliance tasks, but interdependent dimensions of a coherent institutional architecture.

5. Conclusions

The convergence of Explainable AI, digital twin technology, and the EU AI Act marks an inflection point for the credit industry - one that demands not incremental adjustment but a fundamental rethinking of how AI-powered decisions are governed. This paper has argued that standalone ML models and ad-hoc XAI tools, however sophisticated individually, are structurally insufficient to meet the transparency, auditability, fairness, and human oversight requirements that the AI Act now imposes on high-risk credit scoring systems. The proposed Extended Digital Twins of the Credit Deliberation Process (DT-CDP), complemented by the Credit Officer Digital Twin (CODT), represents a shift in conceptual register: from explaining decisions after the fact to governing the entire deliberation process by design.

The framework makes three contributions that extend beyond its specific application domain. Theoretically, it bridges the XAI and digital twin literatures - fields that have developed largely in parallel - by embedding model-level explanation methods within a process-level simulation and governance architecture. This integration responds directly to documented failures of isolated XAI tools: the instability of local surrogate explanations, the misalignment between technical outputs and the needs of borrowers and auditors, and the disconnect between model performance metrics and the broader governance of the institutions deploying those models. Practically, the DT-CDP offers a compliance-by-design path for financial institutions navigating the AI Act's obligations for high-risk systems, transforming regulatory requirements from a checklist exercise into an embedded feature of operational infrastructure. The CODT introduces an additional dimension largely absent from existing governance frameworks: the systematic augmentation of human judgment, surfacing individual biases and inefficiencies not to discipline or replace human decision-makers, but to support the kind of meaningful oversight that the AI Act explicitly requires.

Several limitations must be acknowledged. The framework remains conceptual; empirical validation through implementation in a real banking environment is necessary to assess feasibility, cost, and operational performance. The computational demands of maintaining a real-time digital twin with integrated SHAP and counterfactual generation at scale may be substantial, requiring further research into efficient approximation methods that preserve explanation quality without prohibitive infrastructure costs. The framework also assumes the availability of high-quality, structured data flows from the physical credit process to the digital twin - an assumption that may not hold in institutions with fragmented or legacy IT architectures, where integration costs could prove a significant barrier to adoption.

Future research should pursue four directions. First, empirical pilots of EDT-CDP implementations in banking environments would provide the evidence base needed to move from conceptual proposal to operational recommendation, measuring the impact on explanation quality, bias reduction, and regulatory compliance cost. Second, federated learning approaches merit investigation as a means of enabling digital twins to share insights and calibrate fairness metrics across institutions without transferring sensitive borrower data - a mechanism that could accelerate the development of sector-wide governance standards.

Third, as AI Act implementing regulations evolve, research into standardized explanation templates and fairness metrics aligned with regulatory expectations would reduce compliance uncertainty and enable more consistent cross-institutional comparison. Fourth, the emergence of generative AI methods capable of producing natural-language summaries from SHAP and LIME outputs opens the possibility of dynamically tailoring explanations to the cognitive and informational needs of specific audiences - borrowers, underwriters, and supervisors alike - a direction that could substantially close the gap between technical explainability and the legal standard of "meaningful" explanation.

As the AI Act's compliance timeline for high-risk AI systems approaches its operational threshold in August 2026, financial institutions face a choice: continue treating explainability as an afterthought, patching opaque systems with post-hoc tools, or embed responsible AI governance at the core of credit decisioning. The EDT-CDP framework offers a blueprint for the second path - one in which regulatory compliance is not a burden to be minimized but a source of institutional resilience, decision quality, and ultimately competitive advantage.

6. References

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Boston Consulting Group. (2025). Tech in Banking 2025: How digital twins enhance banking resilience. Retrieved from https://www.linkedin.com/posts/bcg-on-financial-institutions_techinbanking2025-operationalresilience-digitaltwins-activity-7357004186242240512-ksTe

Chatterjee, P. et al. (2024). Digital twin for credit card fraud detection. Future Generation Computer Systems. https://doi.org/10.1016/j.future.2024.04.057

EU AI Act — Regulation (EU) 2024/1689 of the European Parliament and of the Council. Official Journal of the European Union.

Grieves, M. (2014). Digital Twin: Manufacturing Excellence through Virtual Factory Replication. White Paper.

Grieves, M. and Vickers, J. (2017) Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. In: Kahlen, J., Flumerfelt, S. and Alves, A., Eds., Transdisciplinary Perspectives on Complex Systems, Springer, Cham, 85-113. https://doi.org/10.1007/978-3-319-38756-7_4

Hacker, P., & Eber, M. (2025). The Future of Credit Underwriting and Insurance Under the EU AI Act. Harvard Data Science Review.7 (3). https://doi.org/10.1162/99608f92.171157d2

Juliussen, B. A. (2025). The Right to an Explanation Under the GDPR and the AI Act. Lecture Notes in Computer Science (LNCS). https://doi.org/10.1007/978-981-96-2071-5_14

Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30 (NeurIPS 2017). https://doi.org/10.48550/arXiv.1705.07874

Metikoš, Ljubiša & Ausloos, Jef. (2025). The right to an explanation in practice: insights from case law for the GDPR and the AI Act. Law, Innovation and Technology. 17. 1-36. 10.1080/17579961.2025.2469349

Nagaraj, S. K. S. (2025). An Analytical Framework for Bias Mitigation in Credit Scoring Systems through Fairness-Constrained Neural Optimization. International Journal of Artificial Intelligence Data Science and Machine Learning, 6(1), 186–195. https://doi.org/10.63282/3050-9262.IJAIDSML-V6I1P120

Ogbuefi, E., Aifuwa, S. E., Olatunde-Thorpe, J., & Akokodaripon, D. (2025). Explainable AI in Credit Decisioning: Balancing Accuracy and Transparency. International Journal of Advanced Multidisciplinary Research and Studies, 5(5). https://doi.org/10.62225/2583049X.2025.5.5.5024

Pathi, S. P., & Pothineni, J. S. (2025). Interpretable AI in Credit Scoring: A Comparative Survey of SHAP, LIME, and Hybrid Approaches. The American Journal of Engineering and Technology, 7(11), 151–155. https://doi.org/10.37547/tajet/v7i11-304

Pattabhi, A. (2025). Financial Digital Twins: AI and Simulation-Based Risk Management for Banking Systems. International Journal of Artificial Intelligence Data Science and Machine Learning, 6(2), 35–44. https://doi.org/10.63282/3050-9262.IJAIDSML-V6I2P104

Prade, H. (2022). Explainable AI for Transparency in Algorithmic Credit Decisions. Academia Nexus Journal.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144. https://doi.org/10.1145/2939672.2939778

Sowmiya, M. N., Jaya Sri, S., Deepshika, S., & Hanushya Devi, G. (2024). Credit Risk Analysis using Explainable Artificial Intelligence. Journal of Soft Computing Paradigm, 6(3), 272–283. https://doi.org/10.36548/jscp.2024.3.004

Takahashi, D., Shimizu, S., & Tanaka, T. (2024). Counterfactual Explanations of Black-box Machine Learning Models using Causal Discovery with Applications to Credit Rating. arXiv preprint. https://doi.org/10.48550/arXiv.2402.02678

Wachter, S., Mittelstadt, B., & Russell, C. (2018). Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the GDPR. Harvard Journal of Law & Technology, 31(2), 841–887. https://doi.org/10.48550/arXiv.1711.00399