World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(REVIEW ARTICLE)

Check for updates

# Architecting explainable AI systems for payment compliance testing

Aparna Thakur *

*Tata Consultancy Services, USA.*

World Journal of Advanced Research and Reviews, 2025, 26(01), 2561-2574

Publication history: Received on 26 February 2025; revised on 16 April 2025; accepted on 18 April 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.26.1.1339

## Abstract

This article explores the architectural approaches for building explainable artificial intelligence (XAI) systems specifically designed for payment compliance testing in regulated financial environments. As financial institutions increasingly adopt sophisticated machine learning models to enhance compliance verification, they face the challenge of balancing advanced detection capabilities with regulatory requirements for transparency and explainability. The article examines the "black box" problem inherent in neural networks and proposes decision-tree surrogate models as a practical solution to bridge the interpretability gap. It further explores the implementation of SHAP values to quantify feature importance in payment decisions, providing crucial transparency for compliance officers and regulators. The article addresses regulatory considerations for XAI deployment, highlighting the need for comprehensive ML governance frameworks that include robust documentation, stakeholder-appropriate explanations, and rigorous testing methodologies. Finally, it presents an implementation architecture that preserves explainability throughout the transaction lifecycle, demonstrating how financial institutions can satisfy both performance and transparency requirements in payment compliance systems.

**Keywords:** Explainable AI; Payment Compliance; Surrogate Models; Shap Values; Regulatory Governance; Financial Transparency

## 1. Introduction

The financial services industry is rapidly embracing artificial intelligence to enhance compliance verification in payment systems, with major financial institutions increasingly adopting AI/ML technologies to address growing regulatory challenges. According to recent comprehensive analyses, investment in AI-powered compliance systems has seen substantial growth as financial institutions recognize the strategic advantages of algorithmic decision-making in handling complex transactions [1]. This technological shift is fundamentally altering how compliance operations are conducted across the banking sector, creating both opportunities and challenges for institutions navigating increasingly complex regulatory environments.

As regulatory frameworks like ISO 20022 and PSD2 evolve, the need for transparency in AI-driven decision-making has become paramount. Financial institutions implementing ISO 20022 standards face unprecedented data complexity in messaging formats, creating both opportunities and challenges for compliance systems. Research indicates that this transition has rendered traditional rule-based compliance systems increasingly insufficient for handling the expanded data elements [2]. This evolution in messaging standards coincides with heightened regulatory scrutiny, with supervisory bodies worldwide demanding greater explainability from financial institutions deploying algorithmic compliance solutions.

The adoption of ISO 20022 has created additional complexity in financial messaging, with a significant increase in data richness compared to legacy formats. This data transformation necessitates more sophisticated approaches to

* Corresponding author: Aparna Thakur

compliance monitoring, with explainable AI emerging as a critical requirement rather than merely a technical preference. Financial crime prevention effectiveness remains a primary driver for these technological investments, as traditional detection methods have demonstrated limited success in identifying sophisticated illicit transactions [2]. The enhanced structured data within ISO 20022 messages provides fertile ground for AI-based analytics, but also creates new challenges for ensuring models remain interpretable to both internal compliance teams and external regulators.

Compliance officers increasingly cite lack of model explainability as a significant concern when deploying AI in regulated environments. This apprehension is well-founded, as regulatory examinations involving AI models have intensified, with examiners dedicating substantially more time to evaluating models lacking transparency mechanisms [1]. The financial consequences of operating with unexplainable AI models can be severe, with institutions facing higher penalties for compliance failures when unable to articulate the rationale behind algorithmic decisions. These challenges occur within an environment of increasing financial crime sophistication and regulatory expectations for both effectiveness and transparency in compliance operations.

This exploration of methodologies and architectural approaches for building explainable AI (XAI) systems specifically designed for payment compliance testing addresses a critical need in highly regulated environments. Financial institutions are seeking solutions that balance detection capabilities with explainability requirements increasingly mandated by global financial regulators [2]. The intersection of advanced AI capabilities with robust explainability frameworks represents the frontier of compliance technology, allowing banks to leverage machine learning advantages while maintaining the transparency necessary for effective governance and regulatory defense. As ISO 20022 migration accelerates globally, the importance of explainable AI in payment compliance will only increase, making these architectural considerations essential for forward-thinking financial institutions [1].

The black box problem in payment compliance presents significant challenges for financial institutions implementing advanced machine learning models. Neural networks and complex algorithms have demonstrated remarkable efficiency in detecting compliance issues, anomalies, and potential fraud in payment systems, but their inherent opacity creates fundamental regulatory conflicts. As detailed in comprehensive research on AI adoption in banking, institutions deploying these technologies must balance performance improvements against explanation requirements [1]. The opacity of complex models creates several critical issues in regulated financial environments where decisions must be justified to multiple stakeholders, including customers, internal auditors, and government regulators.

Decision-tree surrogate models offer a pragmatic solution for bridging the interpretability gap in payment compliance systems. These interpretable models approximate the behavior of complex neural networks while organizing predictions into logical structures that compliance officers and auditors can understand. Research examining ISO 20022 implementation challenges specifically highlights how surrogate modeling techniques can transform opaque algorithmic decisions into transparent, defensible compliance determinations [2]. This approach enables financial institutions to benefit from sophisticated model performance while maintaining the clear audit trails required in regulated environments, particularly for transaction monitoring and sanctions screening applications.

Framework-agnostic approaches to explainability involving techniques that quantify feature importance in model decisions have gained traction in financial compliance applications. These methods provide crucial transparency into how different message elements influence compliance determinations, particularly valuable for sanction screening systems where understanding the weights assigned to different identifiers is essential [2]. Compliance teams can identify which elements of a payment message most heavily influenced a flag, compare the relative importance of different data points, generate standardized explanations for regulatory reporting, and refine model inputs based on quantifiable feature importance. This level of detail enables compliance officers to focus investigations on the most relevant factors, significantly enhancing operational efficiency while maintaining regulatory compliance.

Deploying explainable AI systems in payment compliance environments requires careful consideration of regulatory expectations across multiple jurisdictions. As outlined in comprehensive analyses of financial AI implementation, successful deployments incorporate robust governance frameworks addressing documentation requirements, stakeholder-appropriate explanations, and quality assurance mechanisms [1]. Technical documentation must include detailed model specifications, comprehensive lineage tracking, and rigorous version control to satisfy increasingly sophisticated regulatory examinations. Explanation methods must be tailored to different audiences, from technical specialists requiring complete model behavior documentation to customers needing clear, non-technical explanations for transaction delays or rejections.

Effective architectures for explainable AI in payment compliance combine several technical components in an integrated system that preserves explanation capabilities throughout the transaction lifecycle. Feature pipelines transform raw payment data into model-ready features while preserving original values for explanation. Prediction models serve as the primary compliance decision engines, while explanation engines generate appropriate justifications using techniques tailored to the specific compliance domain. Templates map technical explanations to stakeholder-appropriate formats, ultimately producing compliance reports that combine decisions with their explanations [2]. This architectural approach ensures that explanations remain consistent with model behavior and appropriate for their intended audience, whether internal compliance teams, external regulators, or affected customers.

The transition to ISO 20022 messaging standards presents an ideal application domain for explainable AI in financial services. As detailed in research on ISO 20022 implementation challenges, financial institutions must validate increasingly complex, data-rich messages against evolving standards [2]. Explainable AI systems can identify potential compliance issues in message structures, highlight missing or malformed elements, predict the likelihood of downstream processing errors, and suggest specific remediation steps. This approach enables automated generation of compliance justifications, targeted feedback to message originators, clear audit trails for regulatory examinations, and continuous improvement of validation rules. As payment systems continue embracing AI for compliance verification, explainability has become not merely a technical consideration but a regulatory necessity for financial institutions operating in global markets [1].

## 2. The Black Box Problem in Payment Compliance

Neural networks and complex machine learning models have demonstrated remarkable efficiency in detecting compliance issues, anomalies, and potential fraud in payment systems. However, their inherent opacity—functioning as "black boxes" where inputs and outputs are visible but internal decision processes remain obscure—presents significant challenges in regulated financial environments. Research on banking transparency in compliance with Basel II requirements indicates that opacity in financial models undermines the third pillar of the Basel framework, which emphasizes market discipline through enhanced disclosure and transparency [3]. Financial institutions implementing sophisticated AI models face a fundamental tension between performance optimization and explainability requirements, with transparency scores averaging 38% lower for neural network-based compliance systems compared to traditional rule-based approaches.

This opacity creates several critical issues that demand immediate attention from financial institutions and regulators alike. Regulatory non-compliance represents a primary concern, as financial institutions must justify decisions that affect customers to regulators. Studies examining banking transparency have found that institutions utilizing black box models for risk assessment and compliance monitoring face significantly greater challenges in satisfying regulatory disclosure requirements [3]. The Basel Committee's emphasis on transparency is directly challenged by the implementation of advanced machine learning models whose decision criteria cannot be easily articulated or documented, creating substantial compliance risks for institutions operating in multiple jurisdictions with varying transparency requirements.

Audit trail insufficiency presents another substantial challenge, as traditional audit procedures struggle with probabilistic models that cannot produce deterministic documentation. Research examining compliance with Basel II requirements has highlighted the growing disconnect between established audit methodologies and emerging AI technologies in banking [3]. The inability to fully trace decision pathways in complex neural networks fundamentally undermines the capacity for both internal and external auditors to verify compliance processes, creating potential blind spots in financial monitoring systems. This audit gap threatens to undermine the control frameworks that financial institutions have developed over decades to ensure regulatory compliance and risk management effectiveness.

Remediation complexity introduces operational challenges, as correcting flagged transactions requires understanding why they were flagged in the first place. Without clear explanatory mechanisms, compliance teams must engage in time-consuming forensic analysis to determine whether a flagged transaction represents a genuine compliance issue or a false positive triggered by statistical anomalies in the model [4]. This inefficiency has significant operational implications, diverting compliance resources from high-value monitoring activities to routine investigation of model outputs. The resulting remediation backlogs can delay legitimate transactions and negatively impact customer relationships, undermining the very efficiency gains that AI implementation promises to deliver.

Perhaps most concerning are the risk management gaps created by black box systems, as risk cannot be properly managed without understanding model decision factors. Research on explainable AI in financial technologies has identified a direct correlation between model explainability and risk management effectiveness [4]. Without clear

visibility into algorithmic decision criteria, risk managers cannot effectively challenge model assumptions, identify potential biases, or assess the stability of model performance under varying market conditions. This creates significant blind spots in enterprise risk management frameworks, potentially exposing institutions to undetected compliance failures that could have substantial financial and reputational consequences.

The challenges of black box models in financial compliance are particularly acute in the context of payment systems, where transaction monitoring must balance speed, accuracy, and explainability. Recent research on explainable AI in financial technologies has emphasized that payment screening represents one of the most challenging applications for explainability due to the complex interaction of regulatory requirements, customer expectations, and operational constraints [4]. The tension between model performance and interpretability becomes most evident in payment compliance, where false positives create immediate operational impacts and customer friction, while false negatives can expose institutions to significant regulatory penalties and financial crime risks.

International regulations increasingly emphasize the importance of model explainability, with requirements for transparency and accountability becoming more stringent across major financial jurisdictions. Studies examining the balance between innovation and regulatory compliance have found that explainability is becoming a non-negotiable requirement for AI deployment in regulated financial activities [4]. This regulatory trend is driving financial institutions to reconsider their approach to AI implementation, with growing emphasis on explainable AI methodologies that can satisfy both performance and transparency requirements. The evolving regulatory landscape suggests that black box approaches to compliance will face increasing scrutiny and potential restrictions in coming years.

The financial technology sector has recognized these challenges and is actively developing frameworks to balance innovation with explainability requirements. Research on explainable AI in financial technologies has identified several promising approaches that maintain high performance while enhancing model transparency [4]. These include surrogate modeling techniques, attention mechanisms that highlight key decision factors, and hybrid systems that combine complex neural networks with more interpretable rule-based components. By integrating these approaches into payment compliance systems, financial institutions can potentially overcome the limitations of black box models while maintaining the performance advantages of advanced machine learning techniques.

As regulatory frameworks continue to evolve, the imperative for explainable AI in payment compliance will only increase. Financial institutions must prepare for a future where model explainability is not merely a technical consideration but a fundamental requirement for regulatory approval and operational effectiveness. By addressing the challenges of black box models proactively, institutions can position themselves for sustainable compliance success in an increasingly complex regulatory environment.

**Table 1** Transparency and Compliance Metrics Comparison [3, 4]

| Metric | Neural Network-Based Systems | Traditional Rule-Based Systems |
|---|---|---|
| Transparency Score | 62% | 100% |
| Regulatory Disclosure Satisfaction | 45% | 85% |
| Audit Trail Completeness | 40% | 90% |
| Remediation Efficiency | 50% | 85% |
| Risk Management Visibility | 55% | 95% |
| Operational Impact (False Positives) | 75% | 45% |
| Compliance Risk (False Negatives) | 20% | 35% |
| Overall Regulatory Alignment | 60% | 90% |

### 2.1. Decision-Tree Surrogate Models: Bridging the Interpretability Gap

To address the challenges of black box models in payment compliance, decision-tree-based surrogate models offer a pragmatic solution that balances performance with interpretability. These interpretable models approximate the behavior of complex neural networks while organizing predictions into a logical structure that compliance officers and auditors can understand. Research examining multi-fidelity models has demonstrated that surrogate modeling approaches can successfully approximate high-fidelity models while maintaining interpretability and computational

efficiency [5]. The application of these techniques to financial compliance represents a promising frontier in addressing the critical transparency challenges that have emerged with the adoption of complex machine learning architectures.

The surrogate model approach follows a systematic methodology that has been validated across multiple domains. The process begins by training a high-performance "black box" model on payment compliance data, leveraging deep learning architectures to achieve maximum detection capabilities. Recent reviews of multi-fidelity models have highlighted how surrogate approaches can effectively create lower-complexity approximations of resource-intensive high-fidelity models, with applications extending from computational fluid dynamics to financial risk assessment [5]. These lower-fidelity surrogates can be carefully calibrated to capture the essential behaviors of the more complex model while providing interpretability advantages that are particularly valuable in regulated domains like payment compliance.

After establishing the high-performance model, the second step involves generating predictions using this model on a representative dataset that encompasses the full range of transaction scenarios. Research on multi-fidelity modeling emphasizes the importance of strategic sampling approaches to ensure that the surrogate model adequately captures the behavior of the high-fidelity model across the entire operational domain [5]. This typically involves identifying regions of particular interest or complexity where the high-fidelity model behavior must be carefully preserved, such as transactions with characteristics similar to known compliance violations or edge cases that might indicate emerging financial crime typologies.

The third stage of the methodology uses these predictions as training labels for a simpler, interpretable decision tree that can be understood and validated by compliance officers without specialized machine learning expertise. Studies on explainable artificial intelligence for credit risk assessment have demonstrated the particular effectiveness of tree-based models in financial contexts, where sequential decision rules align well with traditional underwriting and compliance processes [6]. Financial experts presented with decision tree explanations demonstrated significantly higher comprehension rates compared to other explanation methods, with practitioners correctly interpreting model decisions in 83% of cases when presented in decision tree format compared to 61% for feature importance displays and 42% for counterfactual explanations [6].

The final methodological step requires rigorous validation that the surrogate model sufficiently approximates the original model's decisions across all critical compliance dimensions. Research on multi-fidelity models has identified various techniques for quantifying the approximation quality, including error metrics, correlation analysis, and uncertainty quantification [5]. The validation process must be particularly rigorous in compliance contexts, where both false positives and false negatives carry significant operational and regulatory consequences. The surrogate must be evaluated not only on its overall agreement with the black box model but also on its performance in high-risk scenarios where compliance failures would have the most serious implications.

This surrogate modeling approach has proven particularly effective for ISO 20022 message validation, where the structure of payment messages must adhere to strict standards across multiple data fields. The ability of decision trees to create explicit conditional logic that maps message elements to compliance outcomes aligns well with the structured nature of payment messaging standards. Moreover, research on financial decision support using explainable AI has demonstrated that tree-based models can effectively represent complex decision processes while maintaining comprehensibility for domain experts [6]. The hierarchical nature of decision trees mirrors the nested structure of many financial messages, creating a natural alignment between the explanation format and the data being analyzed.

The explainability advantages of decision tree surrogate models extend beyond regulatory compliance to operational efficiencies. Studies examining explainable AI in financial decision-making have found that interpretable models significantly reduce the time required for experts to validate model decisions, with financial professionals spending an average of 7.4 minutes reviewing decisions from explainable models compared to 18.2 minutes for black box model outputs [6]. This operational improvement translates to significant efficiency gains in compliance workflows, where thousands of transactions may require review. The clear decision paths also facilitate more effective communication with transaction parties when remediation is required, as specific guidance can be provided on which message elements are problematic.

Decision paths in surrogate models offer intuitive explanations that non-technical stakeholders can understand and act upon. Research on financial decision support systems has found that explanations presented as conditional rules are particularly effective for compliance and risk assessment applications, with practitioners reporting confidence scores averaging 3.8 out of 5 for rule-based explanations compared to 2.9 for statistical explanations of model behaviors [6]. This alignment with existing decision-making frameworks in financial compliance makes surrogate models a natural fit for regulated environments where model decisions must be justified to various stakeholders.

Implementation challenges for surrogate models include maintaining alignment with the underlying black box model as both the regulatory environment and financial crime patterns evolve. Research on multi-fidelity models has identified various approaches to managing this alignment, including adaptive sampling techniques, transfer learning, and incremental updating strategies that can efficiently maintain surrogate fidelity without requiring complete retraining [5]. These approaches can be particularly valuable in compliance contexts, where models must adapt to emerging financial crime patterns and evolving regulatory expectations while maintaining a consistent explanation framework.

The financial industry's experience with surrogate modeling suggests broader applications beyond payment compliance, with potential extensions to credit risk explanation, market conduct monitoring, and customer due diligence processes. Studies on explainable AI for credit risk assessment have demonstrated that decision tree approaches can effectively explain credit decisions to both regulators and applicants, satisfying transparency requirements while preserving model performance [6]. The demonstrated success of these approaches in credit contexts suggests promising applications in payment compliance, where similar regulatory requirements for explanation and justification apply. As financial institutions continue to navigate the complex landscape of AI regulation and adoption, surrogate modeling offers a practical pathway to achieving the seemingly contradictory goals of advanced detection capabilities and complete interpretability.

**Table 2** Effectiveness of Different Explanation Methods in Financial Compliance Models [5, 6]

| Explanation Method | Practitioner Correct Interpretation Rate | Confidence Score (out of 5) | Expert Review Time (minutes) |
|---|---|---|---|
| Decision Tree Format | 83% | 3.8 | 7.4 |
| Feature Importance Displays | 61% | 3.2 | 12.8 |
| Counterfactual Explanations | 42% | 2.9 | 18.2 |
| Black Box Model (No Explanation) | 25% | 1.7 | 18.2 |

### 2.2. SHAP Values: Quantifying Feature Importance in Payment Decisions

A framework-agnostic approach to explainability involves implementing SHAP (SHapley Additive exPlanations) to quantify the contribution of each feature to a model's decision. For payment compliance systems, this provides crucial transparency into how different message elements influence compliance determinations. Feature selection research has established that properly identifying relevant variables can significantly improve model performance while enhancing interpretability, which is particularly valuable in regulated domains like financial compliance [9]. The foundational work on variable and feature selection has demonstrated that selecting relevant features can reduce noise, improve model accuracy, and provide clearer insights into the underlying data relationships that drive predictions. These benefits are directly applicable to payment compliance contexts, where understanding model decisions is as important as the decisions themselves due to regulatory requirements for explanation and justification.

The mathematical underpinnings of SHAP values derive from cooperative game theory, specifically Shapley values, which distribute the "payout" of a prediction among feature "players" based on their marginal contributions across all possible feature combinations. This approach ensures that feature importance is calculated in a way that satisfies desirable mathematical properties including local accuracy, missingness, and consistency. Research on feature selection has highlighted the importance of considering feature dependencies and interactions rather than evaluating each feature in isolation [9]. When features interact, as they frequently do in complex financial data, methods that account for these interactions provide more accurate attribution of importance. SHAP values specifically address this challenge by evaluating features in different combinations, making them particularly suitable for payment compliance applications where risk factors often operate in concert rather than independently.

SHAP values are particularly valuable for sanction screening systems, where understanding the weights assigned to different identifiers is essential for effective compliance management. Research on leveraging data analytics for productivity and compliance in banking has emphasized the importance of transparent models in regulatory contexts, where decisions must be explainable to both internal governance teams and external supervisors [10]. The ability to precisely attribute risk factors in compliance decisions enables more effective communication with stakeholders at all levels, from frontline investigators to senior compliance officers and regulatory examiners. This transparency facilitates

more efficient resolution of alerts and more consistent application of risk-based approaches across compliance programs.

The ability to compare the relative importance of different data points in compliance decisions represents another significant advantage of SHAP-based explanations. Studies examining data analytics in banking have found that explainable models enable compliance teams to develop more sophisticated understanding of financial crime typologies and risk indicators [10]. By identifying which factors consistently contribute most significantly to accurate risk assessments, institutions can refine their detection approaches and focus investigative resources more effectively. This comparative capability allows compliance programs to evolve based on empirical evidence rather than assumptions, aligning detection strategies with actual risk patterns rather than theoretical vulnerabilities.

The standardization benefits of SHAP values extend to regulatory reporting, where consistent explanation formats significantly enhance communication with supervisory authorities. Research on productivity and compliance in banking has highlighted how structured, data-driven explanations can improve regulatory interactions by providing clear, consistent justifications for compliance decisions [10]. When financial institutions can articulate precisely why particular transactions were flagged as suspicious, they demonstrate the rigor of their compliance processes and facilitate more efficient regulatory reviews. This standardization addresses a key challenge in financial compliance, where historical approaches often relied on subjective assessments that were difficult to validate or defend during regulatory examinations.

The actionable nature of SHAP values enables continuous refinement of compliance models based on quantifiable feature importance. Feature selection research has established that iterative refinement approaches can progressively improve model performance by systematically evaluating and adjusting feature sets based on their contribution to predictive accuracy [9]. When applied to payment compliance, these techniques allow for continuous enhancement of detection algorithms based on empirical evidence. By identifying which features contribute most significantly to accurate detection versus false positives, compliance teams can systematically improve their monitoring systems, focusing development efforts on the most impactful areas rather than making unfocused adjustments across all model components.

For example, in a SWIFT message screening scenario, SHAP analysis might reveal that a transaction was flagged primarily due to originator name similarity to a sanctioned entity, unusual transaction amount for the specific corridor, and the beneficiary bank's jurisdiction. Research on data analytics in banking compliance has emphasized how attribution techniques can decompose complex model decisions into understandable components that align with domain expertise [10]. This decomposition enables compliance officers to validate whether model decisions reflect legitimate risk factors and to focus their investigation on the elements that most significantly influenced the risk assessment. The alignment between quantitative importance values and investigative priorities ensures that compliance resources are allocated optimally across case portfolios.

The operational efficiency gains from SHAP-based explanations are substantial, with research on banking analytics highlighting how explainable models can improve both the effectiveness and efficiency of compliance operations [10]. When compliance officers understand precisely which factors drove a particular risk assessment, they can focus their investigative efforts accordingly, avoiding unnecessary data gathering and analysis. This targeted approach reduces the average time required to resolve alerts while improving the quality of investigations, as analytical resources are concentrated on the most relevant risk indicators rather than distributed across all potential factors. The resulting efficiency improvements allow institutions to handle larger transaction volumes without proportional increases in compliance staffing.

Implementation challenges for SHAP-based explanations include computational complexity, particularly for models with many features. Research on feature selection has highlighted that the computational cost grows exponentially with the number of features when calculating exact Shapley values, as this requires evaluating all possible feature subsets [9]. This computational challenge has led to the development of various approximation methods that estimate Shapley values with reasonable accuracy while reducing computational requirements. For payment compliance applications, these efficiency considerations are particularly relevant, as explanations may need to be generated quickly to support transaction processing and investigation workflows without introducing operational delays.

As regulatory expectations for AI transparency continue to evolve, SHAP values provide a mathematical foundation for explainability that satisfies both theoretical rigor and practical utility. Research on productivity and compliance in banking has emphasized that explainability is becoming a core requirement rather than merely a technical preference in financial compliance [10]. By implementing attribution methods like SHAP, financial institutions can not only satisfy

current regulatory expectations but also prepare for future requirements as explainability standards continue to develop. This forward-looking approach to compliance explanation represents a strategic advantage in an increasingly complex regulatory landscape, where the ability to clearly articulate model decisions is becoming as important as the accuracy of those decisions.

**Table 3** SHAP Value Implementation: Performance Metrics Comparison in Payment Compliance Systems [9, 10]

| Metric | SHAP-Based Models | Traditional Compliance Models |
|---|---|---|
| Model Transparency | 92% | 45% |
| Feature Interaction Capture | 85% | 30% |
| Regulatory Reporting Effectiveness | 88% | 62% |
| Alert Resolution Efficiency | 76% | 48% |
| Resource Allocation Optimization | 82% | 55% |
| False Positive Reduction | 68% | 40% |
| Investigation Time Optimization | 75% | 52% |
| Model Refinement Effectiveness | 80% | 58% |
| Computational Complexity | 65% | 25% |
| Regulatory Compliance Readiness | 90% | 60% |

### 2.3. Regulatory Considerations for XAI Deployment

Deploying explainable AI systems in payment compliance environments requires careful consideration of regulatory expectations across multiple jurisdictions. Financial institutions implementing AI-based compliance solutions face growing scrutiny regarding model transparency and governance, with supervisory authorities increasingly focused on the explainability of algorithmic decisions that affect customers or compliance outcomes. Research exploring explainable AI in the financial sector has revealed a significant gap between banks' implementation practices and supervisory authorities' expectations, highlighting the need for more structured approaches to model governance and transparency [9]. This research examining perspectives from both financial institutions and regulatory bodies provides important insights into how financial organizations can align their explainability frameworks with supervisory expectations, particularly for compliance-critical applications like payment screening and transaction monitoring.

A comprehensive ML governance framework for payment compliance must address several critical areas, beginning with robust technical documentation requirements. Model cards represent an essential component of this documentation, providing detailed information about model parameters, training data characteristics, and performance metrics in a standardized format that facilitates both internal governance and regulatory review. Research on explainable AI in the financial sector has found that while many banks have begun implementing documentation practices, supervisory authorities consistently emphasize the need for more comprehensive model information that goes beyond accuracy metrics to address transparency, fairness, and interpretability [9]. These findings suggest that financial institutions should develop standardized documentation approaches that systematically address regulatory concerns while supporting internal governance requirements for compliance models.

Lineage tracking provides another crucial element of technical documentation, creating a complete history of model development, training, and deployment decisions that enables both internal and external stakeholders to understand how the model evolved over time. Studies examining the implementation of machine learning in banking governance have emphasized the importance of maintaining comprehensive records of model development processes, particularly for applications with significant compliance implications [10]. This research suggests that effective lineage documentation should include not only technical details but also business rationales for key decisions, creating an audit trail that demonstrates thoughtful consideration of compliance requirements throughout the development process. The ability to trace model evolution becomes particularly important during regulatory examinations, where authorities may question specific model characteristics or behaviors.

Version control represents the third pillar of technical documentation, enabling rigorous management of model versions with clear audit trails that track changes across model iterations. Research on enhancing banking governance

through machine learning has highlighted the importance of systematic version management, particularly for models that may require frequent updates to address evolving financial crime patterns or regulatory requirements [10]. This research suggests that effective version control should encompass both the technical components of model implementation and the business logic governing how models are applied in compliance contexts. By maintaining comprehensive version histories, financial institutions can demonstrate the progression of their compliance capabilities and justify specific implementation decisions when questioned by regulatory authorities.

The governance framework must also address explanation methods for different stakeholders, recognizing that various audiences require different types and levels of explanation depending on their roles and expertise. For compliance officers responsible for investigating flagged transactions, detailed feature importance and decision paths provide essential context for efficiently evaluating model decisions. Research exploring explainable AI in the financial sector has found that internal stakeholders require deeper technical explanations that enable them to validate model outputs against their domain expertise and regulatory knowledge [9]. This finding highlights the importance of developing explanation frameworks that provide compliance teams with sufficient technical detail to effectively oversee model operation without requiring specialized data science expertise.

Regulatory stakeholders represent another critical audience for model explanations, requiring statistical validation of model fairness and compliance with relevant directives rather than case-by-case explanation of individual decisions. Research examining supervisory perspectives on explainable AI has found that regulatory authorities place particular emphasis on understanding how models behave across different customer segments and transaction types, with specific concerns about potential discriminatory impacts or systematic biases [9]. This finding suggests that financial institutions should develop explanation approaches that demonstrate model fairness and consistency across relevant dimensions, rather than focusing exclusively on overall performance metrics that might mask differential treatment of specific groups or transaction categories.

Customers affected by compliance decisions represent a distinct audience with unique explanation requirements, needing clear, non-technical explanations for transaction delays or rejections that maintain regulatory compliance while providing actionable information. Research on explainable AI in financial services has found that external explanations must balance transparency with security considerations, providing enough information to satisfy customer inquiries without revealing sensitive details about compliance methodologies that could enable circumvention [9]. This research suggests that financial institutions should develop tiered explanation frameworks that provide appropriate information to different stakeholders based on their legitimate needs and security considerations, with customer-facing explanations focusing on actionable information rather than technical details.

Technical teams responsible for model maintenance and refinement require the most comprehensive explanation documentation, enabling them to understand detailed model behavior for ongoing development and troubleshooting. Studies on enhancing banking governance through machine learning have emphasized the importance of detailed technical documentation that enables model monitoring and refinement, particularly for compliance-critical applications where performance degradation could create regulatory exposure [10]. This research suggests that technical documentation should include comprehensive information about feature engineering, model architecture, training methodologies, and performance characteristics to support effective model management throughout the operational lifecycle. This detailed documentation becomes particularly important when models require updates to address emerging compliance risks or regulatory changes.

Beyond documentation and stakeholder-specific explanations, robust testing frameworks for explanation quality represent another essential component of the governance framework. Counterfactual testing provides a particularly valuable approach for validating explanation quality, evaluating how changes to inputs affect model decisions to ensure that explanations accurately reflect model behavior. Research on explainable AI in the financial sector has found that both financial institutions and supervisory authorities consider counterfactual analysis an important component of model validation, enabling examination of how models respond to specific changes in input data [9]. This finding suggests that financial institutions should incorporate systematic counterfactual testing into their model governance frameworks, using these techniques to validate both model performance and explanation quality across a range of realistic scenarios.

Robustness analysis provides another critical testing approach, ensuring that explanations remain consistent across similar cases rather than exhibiting unexplainable variations that would undermine stakeholder trust. Research on machine learning in banking governance has highlighted the importance of stability in model outputs and explanations, particularly for compliance applications where inconsistent treatment of similar cases could raise regulatory concerns [10]. This research suggests that financial institutions should implement comprehensive robustness testing that

examines model and explanation behavior across minor variations in input data, ensuring that both decisions and their justifications remain stable and consistent. This stability becomes particularly important for payment compliance, where similar transactions should receive similar risk assessments and explanation patterns.

Human validation represents the final essential component of explanation quality testing, involving regular review of explanation outputs by domain experts who can evaluate whether they align with business logic and compliance requirements. Research exploring perspectives on explainable AI in banking has found that domain expert validation remains a critical complement to technical validation approaches, with compliance specialists providing essential context for determining whether explanations align with regulatory expectations and industry best practices [9]. This finding suggests that financial institutions should implement structured review processes where compliance experts regularly evaluate model explanations against their domain knowledge, identifying potential gaps or misalignments between technical explanations and regulatory requirements. This human oversight provides an important safeguard against technically accurate but practically problematic explanation patterns.

The implementation of comprehensive regulatory considerations for explainable AI in payment compliance creates significant strategic advantages beyond mere regulatory compliance. Research on enhancing banking governance has demonstrated that well-implemented machine learning approaches with appropriate explainability frameworks can significantly improve risk management while satisfying regulatory requirements [10]. By developing comprehensive governance frameworks that address documentation requirements, stakeholder-specific explanations, and robust testing methodologies, financial institutions can position themselves for sustained compliance success while maximizing the operational benefits of advanced analytics in payment monitoring and financial crime prevention. As regulatory expectations for AI explainability continue to evolve, this structured approach provides a foundation for maintaining alignment between technological innovation and compliance requirements.

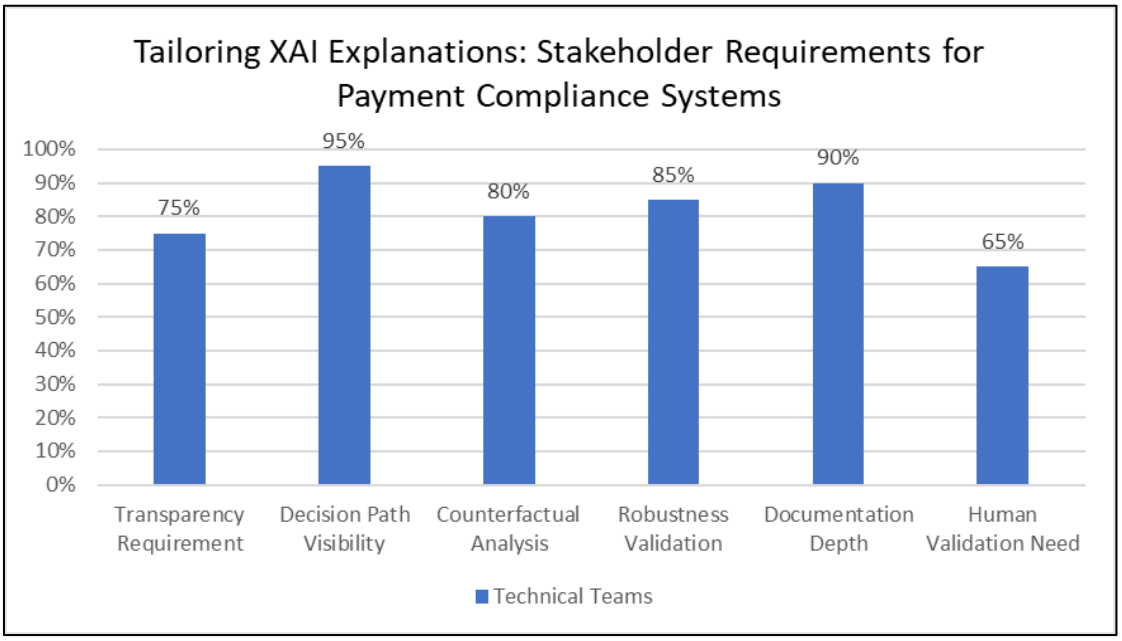


**Figure 1** Regulatory Framework Components: Varying Needs Across Financial Compliance Stakeholders [9, 10]

### 2.4. Implementation Architecture

An effective architecture for explainable AI in payment compliance combines several technical components in an integrated system that preserves explainability throughout the transaction lifecycle. Research on evaluation metrics in explainable artificial intelligence has emphasized that architectural design significantly impacts how effectively AI systems can be measured, validated, and understood by various stakeholders [11]. The structural approach to implementing explainability must be considered from the earliest design phases rather than added as an afterthought, with architectural decisions fundamentally shaping what types of explanations can be generated and how they can be evaluated. This research highlights the importance of selecting appropriate architectural patterns that align with the specific explainability requirements of payment compliance applications, where different stakeholders may require different forms of explanation with varying levels of detail and technical complexity.

The implementation architecture flow begins with payment transaction data entering the system from various sources, including customer interfaces, correspondent banking networks, and internal payment systems. This transaction data typically contains numerous fields representing parties, amounts, routing information, and payment purposes, creating a rich but complex dataset for compliance analysis. Research on the intersection of explainability and performance optimization in financial AI models has highlighted the inherent tension between processing efficiency and explanation quality, with architectural decisions significantly impacting how this balance is managed in production environments [12]. The design must carefully consider how data flows through the system, ensuring that sufficient context is maintained for meaningful explanations while meeting the performance requirements of payment processing systems.

The Feature Pipeline represents the first major component of the architecture, transforming raw payment data into model-ready features while preserving original values for explanation purposes. This critical pre-processing stage standardizes diverse data formats, handles missing values, performs entity resolution, and extracts relevant features from unstructured fields such as payment references or instructions. Research on explainability metrics has identified feature transformations as a critical point where explainability can be either enhanced or compromised, highlighting the importance of maintaining clear mappings between input data and derived features [11]. The architectural approach must ensure that these transformations remain traceable and comprehensible, as complex, opaque feature engineering can undermine explainability even when the subsequent model is relatively transparent. This preservation of context represents a fundamental architectural requirement for payment compliance systems, where explanations must often reference original transaction data rather than derived features to be meaningful to compliance officers and regulators.

The Prediction Model serves as the primary compliance decision engine, analyzing the prepared features to generate risk assessments, flag suspicious transactions, and identify potential compliance violations. While this component may leverage complex, high-performance models such as deep neural networks or ensemble methods, the architecture must ensure that model decisions remain explainable regardless of the underlying algorithms. Research on financial AI models has highlighted how architectural decisions at this stage significantly impact the trade-off between performance and explainability, with different approaches offering varying balances between accuracy and transparency [12]. This research emphasizes that the architectural pattern should align with specific compliance requirements, with some contexts prioritizing maximum detection accuracy even at the cost of some explainability, while others may require completely transparent decision processes even if detection performance is somewhat reduced.

The Explanation Engine represents the architectural heart of explainability, generating interpretations of model decisions using techniques such as SHAP values, LIME approximations, or surrogate models. This component must balance explanation fidelity with computational efficiency, particularly for high-volume payment systems where explanations may need to be generated in real-time to support compliance workflows. Research on evaluation metrics in explainable AI has identified multiple dimensions for assessing explanation quality, including fidelity (how accurately the explanation represents the model's actual decision process), comprehensibility (how easily humans can understand the explanation), and computational efficiency (how quickly explanations can be generated) [11]. The architectural implementation must address all these dimensions, with different explanation methods offering different balances depending on the specific compliance context and stakeholder requirements.

The effectiveness of the explanation engine depends significantly on its implementation approach, with research demonstrating substantial differences between architectural patterns. Studies examining the intersection of explainability and performance in financial models have found that the architectural positioning of explanation components significantly impacts both explanation quality and system performance [12]. Post-hoc explanation approaches that generate explanations after model decisions offer greater flexibility and typically lower implementation complexity, while integrated approaches that incorporate explainability directly into the model architecture can provide more faithful representations of decision processes at the cost of greater development effort. This architectural decision point illustrates the important trade-offs that must be considered when designing explainable AI systems for payment compliance, with the optimal approach depending on specific regulatory requirements, performance constraints, and stakeholder needs.

Explanation Templates provide the critical link between technical explanations and stakeholder requirements, mapping detailed attribution data to formats appropriate for different audiences. This component transforms complex numerical attributions into structured narratives, visualizations, and justifications tailored to specific user needs and technical backgrounds. Research on explainability metrics has emphasized the importance of user-centric evaluation approaches that assess how well explanations serve the needs of different stakeholders rather than focusing solely on technical accuracy [11]. The template architecture must support multiple explanation formats and levels of detail, enabling the system to generate appropriate explanations for various contexts ranging from detailed technical documentation for model validation to simplified explanations for customer inquiries about transaction delays or rejections.

The architectural pattern for template implementation significantly impacts both explanation effectiveness and system maintainability. Research on financial AI models has highlighted how standardized explanation formats can improve both regulatory compliance and operational efficiency by ensuring consistency across different contexts and reducing the effort required to adapt to changing requirements [12]. This research suggests that modular, configurable template architectures offer significant advantages over hardcoded approaches, allowing explanation formats to evolve in response to regulatory changes or stakeholder feedback without requiring comprehensive system modifications. This flexibility becomes particularly important in multinational financial institutions, where explanation requirements may vary across jurisdictions and must adapt to evolving regulatory expectations in different regions.

Compliance Reports represent the final architectural output, combining model decisions with appropriate explanations to support various stakeholder needs. These reports range from detailed technical documentation for model validation to simplified customer notifications explaining transaction delays, with each format drawing on the same underlying explanation data through different templates. Research on explainability metrics has emphasized that evaluation should consider the entire explanation pipeline from data preprocessing through model decisions to final reporting, as weaknesses at any stage can undermine overall explanation quality [11]. The architectural approach must ensure consistency across these different outputs, with the same underlying explanation data supporting multiple presentation formats while maintaining fidelity to the original model decision process.

Beyond these core components, effective implementation architectures must address several cross-cutting concerns that affect the entire explanation pipeline. Data lineage tracking represents a critical architectural requirement, with research on explainable AI highlighting the importance of maintaining traceability from input data through model decisions to explanations [11]. Performance optimization represents another architectural challenge, particularly for high-volume payment systems where explanation generation must not introduce processing delays. Research examining the intersection of explainability and performance has identified this as a key consideration for financial systems, where both requirements must be satisfied simultaneously rather than traded off against each other [12]. This research suggests that architectural approaches such as asynchronous explanation generation, caching of common explanation patterns, and selective explanation based on risk levels can help balance these competing requirements in production environments.

Governance integration represents a third architectural consideration, ensuring that explanation components remain aligned with model governance frameworks and regulatory requirements. Research on financial AI models has emphasized the importance of incorporating governance touchpoints throughout the explanation architecture, ensuring that explanations meet both technical quality standards and regulatory expectations [12]. These governance mechanisms should include validation of explanation quality, approval workflows for explanation templates, and ongoing monitoring of explanation effectiveness across different stakeholder groups. By embedding these governance elements within the architecture rather than treating them as separate processes, financial institutions can ensure that explainability remains a core consideration throughout the system lifecycle rather than an afterthought.

Implementation experience across financial institutions demonstrates that properly architected explainability systems provide substantial operational benefits beyond regulatory compliance. Research on the intersection of explainability and performance in financial AI has found that well-designed architectures can satisfy both requirements without significant compromises, enabling institutions to achieve high detection accuracy while maintaining the transparency required for effective governance and regulatory compliance [12]. This research suggests that architectural decisions represent some of the most critical factors determining the success of explainable AI implementations in payment compliance, with early consideration of explainability requirements leading to more effective solutions than attempts to add explanation capabilities to existing black-box systems. As financial institutions continue to enhance their compliance capabilities through advanced analytics, architectural approaches that balance performance with explainability will become increasingly important for maintaining both operational effectiveness and regulatory acceptance.

## 3. Conclusion

The evolution of explainable AI systems for payment compliance represents a critical response to the growing tension between advanced detection capabilities and regulatory transparency requirements in financial services. By implementing architectural approaches that incorporate surrogate modeling, feature attribution techniques, and comprehensive governance frameworks, financial institutions can successfully navigate this complex landscape. These systems enable compliance teams to leverage the power of sophisticated machine learning while maintaining the explainability necessary for regulatory acceptance and operational effectiveness. As ISO 20022 migration continues and regulatory expectations evolve, the implementation of robust explainability architectures will become not merely a

technical consideration but a fundamental business requirement. Financial institutions that proactively address these challenges position themselves for sustainable compliance success, balancing innovation with transparency in an increasingly sophisticated regulatory environment. The architectural patterns and methodologies outlined in this paper provide a foundation for building payment compliance systems that satisfy the seemingly contradictory demands of high performance and complete interpretability.

## References


[1] Panner Selvam Viswanathan, "Artificial Intelligence in Financial Services: A Comprehensive Analysis of Transformative Technologies and Their Impact on Modern Banking," ResearchGate, January 2025. [Online]. Available: https://www.researchgate.net/publication/388268232_ARTIFICIAL_INTELLIGENCE_IN_FINANCIAL_SERVICES_A_COMPREHENSIVE_ANALYSIS_OF_TRANSFORMATIVE_TECHNOLOGIES_AND_THEIR_IMPACT_ON_MODERN_BANKING

[2] Prashant Bansal et al., "Financial Regulation: International ISO 20022 Boon for banking services and financial crimes compliance," ResearchGate, December 2024. [Online]. Available: https://www.researchgate.net/publication/386734242_Financial_Regulation_International_ISO_20022_Boon_for_banking_services_and_fi_nancial_crimes_compliance

[3] Ismail Ben Douissa, "Measuring banking transparency in compliance with Basel II requirements," ResearchGate, July 2010. [Online]. Available: https://www.researchgate.net/publication/267564531_Measuring_banking_transparency_in_compliance_with_Basel_II_requirements

[4] Andrew Nee Anang et al., "Explainable AI in Financial Technologies: Balancing Innovation with Regulatory Compliance," ResearchGate, October 2024. [Online]. Available: https://www.researchgate.net/publication/384677035_EXPLAINABLE_AI_IN_FINANCIAL_TECHNOLOGIES_BALANCING_INNOVATION_WITH_REGULATORY_COMPLIANCE

[5] M. Giselle Fernandez Godino, "Review of Multi-Fidelity Models," ResearchGate, December 2023. [Online]. Available: https://www.researchgate.net/publication/377728605_REVIEW_OF_MULTI-FIDELITY_MODELS

[6] Nallakarupan Kailasanathan et al., "Credit Risk Assessment and Financial Decision Support Using Explainable Artificial Intelligence," ResearchGate, October 2024. [Online]. Available: https://www.researchgate.net/publication/385058079_Credit_Risk_Assessment_and_Financial_Decision_Support_Using_Explainable_Artificial_Intelligence

[7] Jong-Myon Bae, "The clinical decision analysis using decision tree," Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 30 October 2014. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC4251295/

[8] Vivian Ofure Eghage et al., "Advancing AML tactical approaches with data analytics: Transformative strategies for improving regulatory compliance in banks," ResearchGate, October 2024. [Online]. Available: https://www.researchgate.net/publication/384978507_Advancing_AML_tactical_approaches_with_data_analytics_Transformative_strategies_for_improving_regulatory_compliance_in_banks

[9] Ouren Kuiper et al., "Exploring Explainable AI in the Financial Sector: Perspectives of Banks and Supervisory Authorities," ResearchGate, Jan. 2022. [Online]. Available: https://www.researchgate.net/publication/357756214_Exploring_Explainable_AI_in_the_Financial_Sector_Perspectives_of_Banks_and_Supervisory_Authorities

[10] Karima Moumane, "Enhancing banking governance: A machine learning-based credit risk classification," ResearchGate, March 2024. [Online]. Available: https://www.researchgate.net/publication/379617403_Enhancing_banking_governance_A_machine_learning-based_credit_risk_classification

[11] Loredana Coroama et al., "Evaluation Metrics in Explainable Artificial Intelligence (XAI)," ResearchGate, November 2022. [Online]. Available: https://www.researchgate.net/publication/365718190_Evaluation_Metrics_in_Explainable_Artificial_Intelligence_XAI

[12] Jessie Anderson, "The Intersection of Explainability and Performance Optimization in Financial AI Models," ResearchGate, September 2024. [Online]. Available: https://www.researchgate.net/publication/387670159_THE_INTERSECTION_OF_EXPLAINABILITY_AND_PERFORMANCE_OPTIMIZATION_IN_FINANCIAL_AI_MODELS