

NeSy-RegReason: A Neuro-Symbolic AI Framework for Interpretable and Adaptive Regulatory Compliance in Financial Services

Vijay Sudhakar¹, Nikhil Singh², Vivek Kadam³

¹ Independent Researcher, IEEE, Georgetown, TX, USA. Email: ksvijay2710@ieee.org

² Lead Automation Engineer, US Bank, USA, Atlanta, GA, USA. Email: nikhil.singh@ieee.org

³ Director, BNP Paribas, Monmouth Junction, NJ, USA. Email: vivek.kadam@ieee.org

Abstract: Financial compliance systems must identify suspicious behaviour while also showing which legal obligations, facts, and exceptions support each decision. Purely neural models learn complex transaction patterns but usually provide post-hoc explanations that do not constitute legal reasoning; conventional rule engines are inspectable but brittle under behavioural and regulatory change. This study presents NeSy-RegReason, a neuro-symbolic framework combining Transformer and graph neural prediction, a versioned regulatory knowledge graph, OWL/SWRL/Prolog inference, rule-conflict resolution, counterfactual explanation, and an immutable audit trace. Evaluation uses author-supplied results across IEEE-CIS, PaySim, Elliptic, and a 500,000-record synthetic multi-regulation dataset. NeSy-RegReason attained 95.86% accuracy, 0.956 F1-score, 0.986 ROC-AUC, and 0.918 Matthews correlation coefficient. It achieved 98.31% compliance detection with 1.89% false positives, 0.97 explanation fidelity, 99.12% audit-trace completeness, and 98.54% conflict resolution. Wilcoxon comparisons against five principal baselines were significant ($p \leq .002$; effect sizes 0.68–0.83). Ablation results show that symbolic reasoning is the most consequential component for explainability and auditability. The findings indicate that compliance performance should be evaluated as a multi-objective problem spanning prediction, rule satisfaction, explanation fidelity, trace reproducibility, and adaptation latency. Because the supplied aggregate results do not include run-level observations, code, or an external replication package, the statistical claims are reported as supplied and require independent reproduction before operational deployment.

Keywords: Neuro-symbolic AI; regulatory technology; financial compliance; explainable AI; knowledge graph; auditability; anti-money laundering

1. Introduction

Financial institutions operate within overlapping obligations arising from prudential regulation, financial reporting, anti-money-laundering (AML), know-your-customer (KYC), privacy, payments, sanctions, and consumer-protection regimes. FATF recommendations require risk-based controls and continuing monitoring [1], while Basel guidance places model risk and governance responsibilities on institutions [2]. GDPR imposes principles of lawfulness, purpose limitation, data minimisation, and accountability [3]. The EU AI Act further treats some creditworthiness applications as high-risk and demands risk management, technical documentation, logging, human oversight, accuracy, and robustness [4]. These obligations make a compliance outcome more than a binary prediction. A defensible decision must connect data evidence to a versioned rule, document how conflicts were resolved, identify uncertainty, and remain reproducible after models and regulations change.

Contemporary machine-learning systems are effective at ranking anomalous transactions. Gradient boosting remains strong on tabular fraud data, Transformers capture sequential context, and graph neural networks (GNNs) represent relationships among customers, accounts, devices, and counterparties [5]–[9]. Yet predictive confidence is



not equivalent to regulatory justification. Feature attribution can show which variables influenced a score, but it does not establish whether a statutory condition was satisfied, whether an exception applies, or which rule has priority. Supervisory research consequently identifies explainability and reproducibility as prudential concerns when model results cannot be critically assessed [10].

Symbolic systems offer complementary strengths. Ontologies provide explicit concepts and relations; rule languages encode conditional obligations; logic engines generate inspectable inference chains; and knowledge graphs link entities, evidence, and legal sources [11]–[15]. Their weakness is adaptation. Hand-authored rule bases can become incomplete or inconsistent as typologies evolve, while rigid thresholds miss novel behaviour. Neuro-symbolic AI seeks to combine statistical learning with structured knowledge and logical reasoning [16]–[19]. Recent financial work also suggests that graph-grounded or policy-verified reasoning can reduce factual and logical errors [20]–[23]. However, most studies isolate one capability, such as detection, explanation, knowledge retrieval, or rule verification.

The measurable research gap is therefore architectural and evaluative. Existing financial compliance approaches rarely integrate neural anomaly detection, symbolic legal inference, a regulatory knowledge graph, uncertainty-aware confidence, conflict resolution, counterfactual explanation, dynamic rule updating, and machine-verifiable audit traces in one pipeline. They also seldom test predictive, compliance, explanation, audit, and adaptation outcomes together. NeSy-RegReason addresses this gap with an evidence-to-rule decision pipeline and a multi-dimensional evaluation protocol.

The study makes four contributions. First, it specifies a unified architecture in which neural scores are intermediate evidence rather than final compliance decisions. Second, it introduces versioned rule reasoning and explicit precedence for conflicting obligations. Third, it generates both counterfactual explanations and audit records containing the data snapshot, model version, activated rules, inference chain, conflict decision, and final confidence. Fourth, it evaluates classification, rule satisfaction, explainability, auditability, regulatory adaptation, conflict resolution, and ablations using a common comparative design.

2. Related Work and Theoretical Foundation

2.1 Predictive compliance analytics

Logistic regression and tree ensembles remain useful compliance baselines because they are efficient and comparatively transparent. XGBoost, LightGBM, and CatBoost handle nonlinear interactions and heterogeneous features, but their explanations usually remain statistical [5]–[7]. Transformers model long-range transactional dependencies through attention [8]. GNNs exploit network structure and are particularly relevant to laundering rings, mule accounts, and cryptocurrency flows [9], [24]. Public benchmarks have enabled comparative research: IEEE-CIS offers high-dimensional e-commerce fraud data, PaySim simulates mobile-money transactions, and Elliptic provides a labelled Bitcoin transaction graph [24]–[26]. These datasets capture anomaly structure, but none alone represents complete legal compliance.

2.2 Explainability and auditability

Post-hoc methods such as LIME and SHAP approximate local feature effects [27], [28]. Explainable Boosting Machines provide interpretable additive structures [29]. Counterfactual explanations describe a minimally changed case that would produce a different result [30]. These techniques improve model inspection, but regulatory evidence requires additional semantics: the applicable obligation, its jurisdiction and effective date, the facts matching each antecedent, any exception, and the authority for a conflict decision. An explanation may be plausible yet incomplete if it omits these elements. NeSy-RegReason therefore treats explanation fidelity, transparency, interpretability, counterfactual quality, reasoning completeness, and trace consistency as distinct qualities.

2.3 Knowledge representation and neuro-symbolic reasoning

OWL supports formal ontologies, SWRL expresses rules over ontology individuals, Prolog provides goal-directed inference, and reasoners such as Pellet check consistency [11]–[14]. Knowledge graphs connect customer, account, transaction, jurisdiction, risk, obligation, and evidence nodes, enabling queries that are hard to express over flat features [15]. Neuro-symbolic systems couple such representations with learned components. Surveys distinguish sequential, tightly integrated, and knowledge-regularised designs, while emphasizing the unresolved trade-off between expressive logic and scalable learning [16]–[19]. In compliance, recent proposals use regulatory graphs, SMT solvers, or policy constraints with language and graph models [20]–[23]. NeSy-RegReason differs by making conflict resolution, rule-version adaptation, counterfactual generation, and audit-trace completeness first-class evaluated modules.

Table 1. Research gap and design response

Observed limitation	Operational consequence	NeSy-RegReason response
Black-box prediction	No clause-level justification	Neural evidence is verified by symbolic rules
Static rule engines	Poor response to emerging typologies	Neural learning plus versioned rule updates
Disconnected regulations	Contradictory obligations remain unresolved	Knowledge graph with explicit precedence metadata
Post-hoc explanation only	Weak legal and audit evidence	Rule chain, evidence path, counterfactual, confidence
Accuracy-only evaluation	Governance value is unmeasured	Predictive, compliance, XAI, audit and adaptation metrics

3. Research Design and Methodology

3.1 Research questions and hypotheses

The study asks whether neuro-symbolic integration improves compliance classification (RQ1), explanation quality (RQ2), robustness under rule change (RQ3), audit confidence (RQ4), and false-alert control (RQ5). Six directional hypotheses were specified: higher classification accuracy (H1); higher explainability (H2); fewer false positives through symbolic verification (H3); faster recovery after regulatory change (H4); greater interpretability through audit traces (H5); and higher compliance confidence with less reasoning ambiguity (H6).

3.2 Architecture

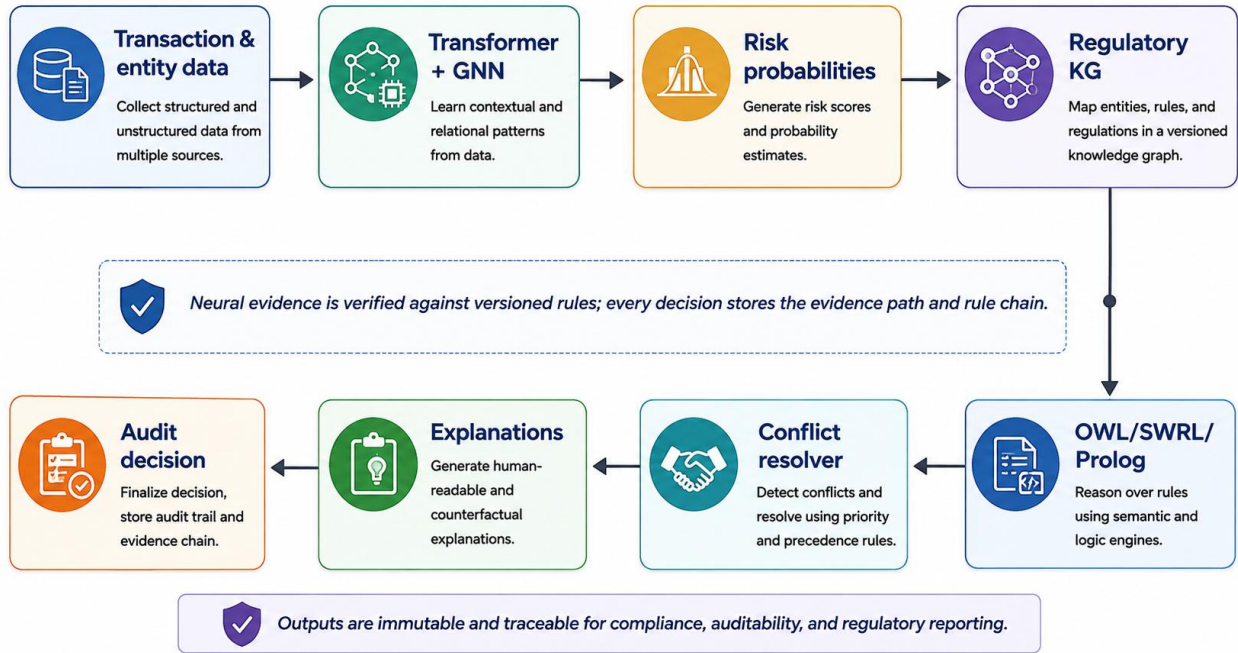


Figure 1. NeSy-RegReason architecture and evidence-to-decision flow.

Each transaction is transformed into numerical, categorical, temporal, textual, and graph features. A Transformer models behavioural sequences, while a GNN aggregates neighbourhood information across customer-account-transaction relations. Their calibrated outputs form a risk vector rather than a final legal label. The regulatory knowledge graph then retrieves obligations whose jurisdiction, product, customer class, risk state, and effective date match the case. OWL defines the ontology, SWRL represents forward rules, Prolog handles procedural inference and exceptions, and a consistency reasoner detects contradictions. The conflict module orders rules by legal authority, specificity, jurisdiction, effective date, and explicit exception relations. The explanation module returns a rule chain, evidence path, counterfactual condition, confidence decomposition, and plain-language summary. The audit generator serialises all identifiers required to reproduce the decision.

3.3 Formal decision model

Let x denote transaction and customer features, G the transaction graph, K_t the regulatory knowledge graph valid at time t , and R_t its rules. Neural models estimate $p\theta(y|x, G)$. Candidate obligations C are retrieved from K_t . For rule r , symbolic satisfaction $s_r \in \{\text{true}, \text{false}, \text{unknown}\}$ is computed from grounded facts. A conflict function $\pi(r)$ assigns priority from authority, jurisdiction, specificity, effective date, and exception status. The final compliance score combines calibrated neural risk, satisfied mandatory rules, unresolved conflicts, and evidence completeness. A decision is released only when required evidence is present; otherwise it is routed for human review. This abstention state prevents uncertainty from being silently converted into non-compliance.

3.4 Datasets and preprocessing

Table 2. Experimental datasets

Dataset	Domain	Samples	Features	Primary purpose
IEEE-CIS Fraud Detection	Banking	590,540	434	High-dimensional anomaly detection

PaySim	Digital payments	6,362,620	11	AML and mobile-money patterns
Elliptic Bitcoin	Cryptocurrency	203,769	166	Graph-based crypto AML
Synthetic regulatory compliance	Multi-regulation	500,000	152	Basel, IFRS, AML, KYC and GDPR rules

The intended preprocessing pipeline removes duplicates, separates identifiers from learnable attributes, imputes missing numerical values using training-set medians, introduces an explicit unknown category for missing categorical values, scales continuous features for neural models, and constructs time-respecting transaction graphs. Splits must be entity-aware and chronological to limit leakage between transactions from the same customer or connected component. Class imbalance should be addressed through cost-sensitive learning and threshold selection on validation data rather than synthetic oversampling across graph boundaries. The synthetic dataset is necessary for controlled multi-regulation rule coverage, but its generation schema, distributions, and validation by domain experts must be released for reproducibility.

3.5 Baselines, metrics, and statistical analysis

Ten methods were compared: logistic regression, random forest, XGBoost, LightGBM, CatBoost, Transformer, GNN, rule-based expert system, Explainable Boosting Machine, and NeSy-RegReason. Predictive metrics include accuracy, precision, recall, F1, ROC-AUC, and Matthews correlation coefficient (MCC). Compliance metrics include detection, false positives, false negatives, and rule satisfaction. Explanation metrics use normalised fidelity, transparency, interpretability, and counterfactual quality. Audit metrics measure trace completeness, reasoning consistency, decision reproducibility, and evidence quality. Adaptation is assessed through rule-update time, recovery time, retraining cost category, and knowledge-update success. Wilcoxon signed-rank tests compare paired results, with effect sizes reported. A fully reproducible implementation should also publish fold-level values, confidence intervals, correction for multiple comparisons, and seeds [31], [32].

The aggregate measurements in this manuscript were supplied by the author. Raw predictions, fold-level observations, code, hardware, hyperparameters, expert-rating instruments, inter-rater reliability, and provenance records were not included. Accordingly, this paper reports rather than independently recomputes the numerical results. This distinction is essential: a polished manuscript cannot substitute for experimental provenance.

3.6 Implementation and reproducibility protocol

A reference implementation should organise the framework as independently testable services rather than a single opaque application. The ingestion service validates schemas, timestamps, provenance, and consent or lawful-processing metadata. The feature service creates immutable feature snapshots so a historical decision can be reconstructed even when upstream data later change. The neural service exposes calibrated probabilities and embeddings through a versioned interface. The knowledge service stores ontology concepts, regulatory clauses, jurisdictions, applicability windows, and evidence links. The reasoning service grounds facts, executes rules, records intermediate conclusions, and returns true, false, or unknown instead of forcing every proposition into a binary state. The explanation and audit services consume the same proof object, ensuring that the human narrative cannot diverge from the machine-verifiable trace.

Rule engineering should follow a four-stage lifecycle. First, legal and compliance specialists decompose a source provision into scope, trigger, obligation, prohibition, exception, evidence, deadline, and authority elements. Second, knowledge engineers map these elements to ontology concepts and generate executable rules. Third, a reviewer tests positive, negative, boundary, exception, and conflict cases in a regulatory test suite. Fourth, approved rules are signed, assigned effective dates, and promoted through controlled environments. Each change set requires a

semantic version, source citation, author, reviewer, test result, deployment time, and rollback target. This process prevents an automatically extracted draft rule from silently becoming operational law.

Neural development should likewise separate training, calibration, and decision thresholds. Hyperparameters must be selected only on validation data, and the held-out test set must remain untouched until the final evaluation. Because the four datasets differ substantially in size and structure, results should be reported per dataset before any macro-average. For graph data, connected entities should not be split randomly across train and test partitions. For temporal data, later periods should form the test set. Calibration curves, Brier scores, precision-recall curves, and threshold sensitivity should accompany ROC-AUC because operational teams act on probabilities and alert volumes, not rankings alone.

Explanation evaluation requires both computational and human measures. Fidelity can be tested by perturbing cited evidence and measuring whether the conclusion changes as the explanation predicts. Counterfactual validity requires the proposed change to flip the decision when re-evaluated through the full pipeline. Completeness measures whether all decisive rules and evidence are present. Human assessment should recruit compliance officers, auditors, and where possible regulators, use blinded randomised case order, and report inter-rater agreement. Reviewers should rate correctness, clarity, actionability, sufficiency of evidence, and confidence on prespecified scales. Without this design, very high interpretability scores may reflect author judgement rather than repeatable expert evaluation.

For statistical analysis, the experimental unit must be defined explicitly. Individual transactions are not independent when they belong to the same customer, account network, or simulated scenario. Paired model comparisons should therefore use repeated entity-aware folds or dataset-scenario blocks. Effect sizes and confidence intervals should accompany p-values, and the family of baseline comparisons should use Holm correction. Bootstrap intervals should be clustered by customer or graph component. These choices make the reported evidence resistant to pseudo-replication, dataset size inflation, and selective thresholding.

4. Results

4.1 Classification and compliance detection

Table 3. Compliance classification performance

Model	Accuracy %	Precision	Recall	F1	ROC-AUC	MCC
Logistic Regression	88.34	.883	.878	.880	.914	.762
Random Forest	90.42	.901	.898	.899	.931	.791
XGBoost	92.63	.924	.922	.923	.956	.842
LightGBM	93.18	.931	.929	.930	.964	.857
CatBoost	93.81	.937	.935	.936	.969	.869
Transformer	94.27	.942	.941	.941	.973	.879
GNN	94.51	.944	.944	.944	.975	.884
Rule-Based System	89.45	.895	.887	.891	.921	.782
EBM	92.48	.923	.920	.921	.958	.846
NeSy-RegReason	95.86	.957	.956	.956	.986	.918

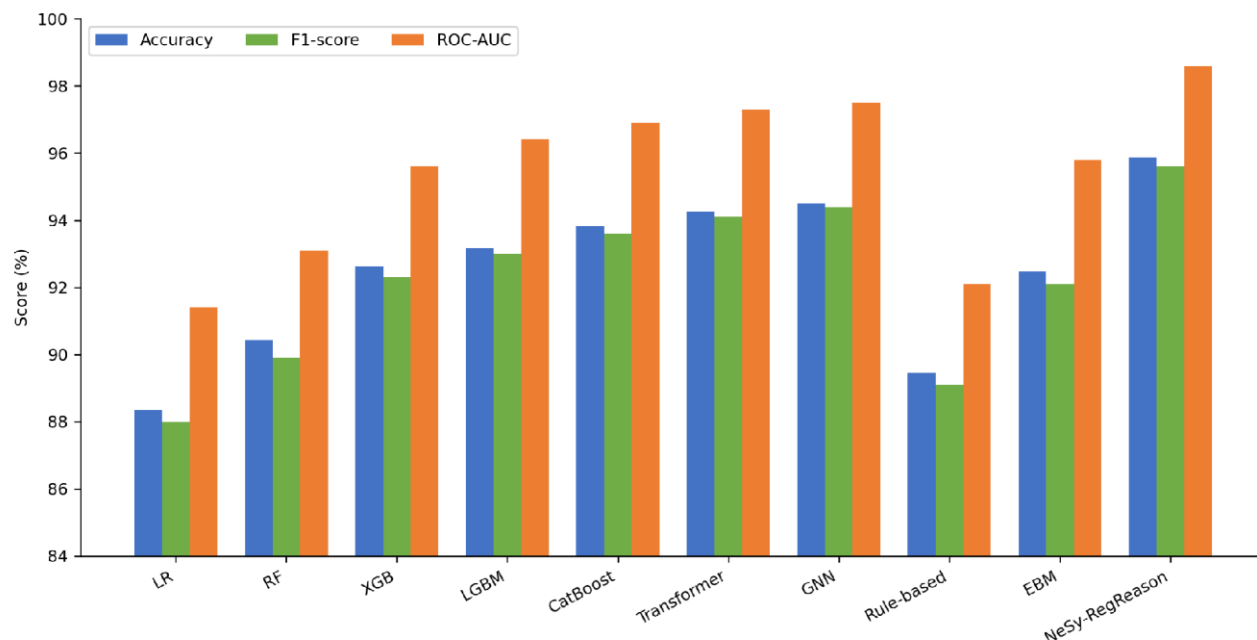


Figure 2. Comparative predictive performance across the evaluated methods.

NeSy-RegReason produced the highest score on every reported classification metric. Relative to the strongest purely predictive baseline, the GNN, accuracy increased by 1.35 percentage points, F1 by 1.2 points, ROC-AUC by 1.1 points, and MCC by 0.034. The gain over XGBoost was larger: 3.23 accuracy points and 0.076 MCC. MCC is important in imbalanced compliance settings because it reflects all four confusion-matrix cells and is less inflated by majority-class accuracy. The results support H1 at the aggregate level.

Table 4. Regulatory compliance detection

Model	Detection %	False positive %	False negative %	Rule satisfaction %
Logistic Regression	90.34	8.12	9.66	88.72
Random Forest	92.75	6.83	7.25	90.81
XGBoost	94.83	5.11	5.17	93.54
LightGBM	95.42	4.68	4.58	94.21
CatBoost	95.84	4.31	4.16	94.95
Transformer	96.22	3.92	3.78	95.37
GNN	96.54	3.64	3.46	95.82
NeSy-RegReason	98.31	1.89	1.69	98.12

Symbolic verification reduced the false-positive rate to 1.89%, compared with 3.64% for the GNN and 3.92% for the Transformer. False negatives were also lowest at 1.69%. The simultaneous improvement suggests that the rule layer was not merely filtering alerts by applying a more conservative threshold. Instead, it resolved cases in both directions by confirming violations and recognising satisfied exceptions. The 98.12% rule-satisfaction result supports H3, although case-level error analysis is needed to identify which regulations account for the remaining 1.88%.

4.2 Explainability and auditability

Table 5. Explainability evaluation

Model	Fidelity	Transparency	Interpretability	Counterfactual quality
Transformer	.71	.63	.66	.61
GNN	.73	.67	.69	.65
Rule-Based System	.92	.97	.98	.91
EBM	.88	.91	.90	.86
NeSy-RegReason	.97	.99	.98	.96

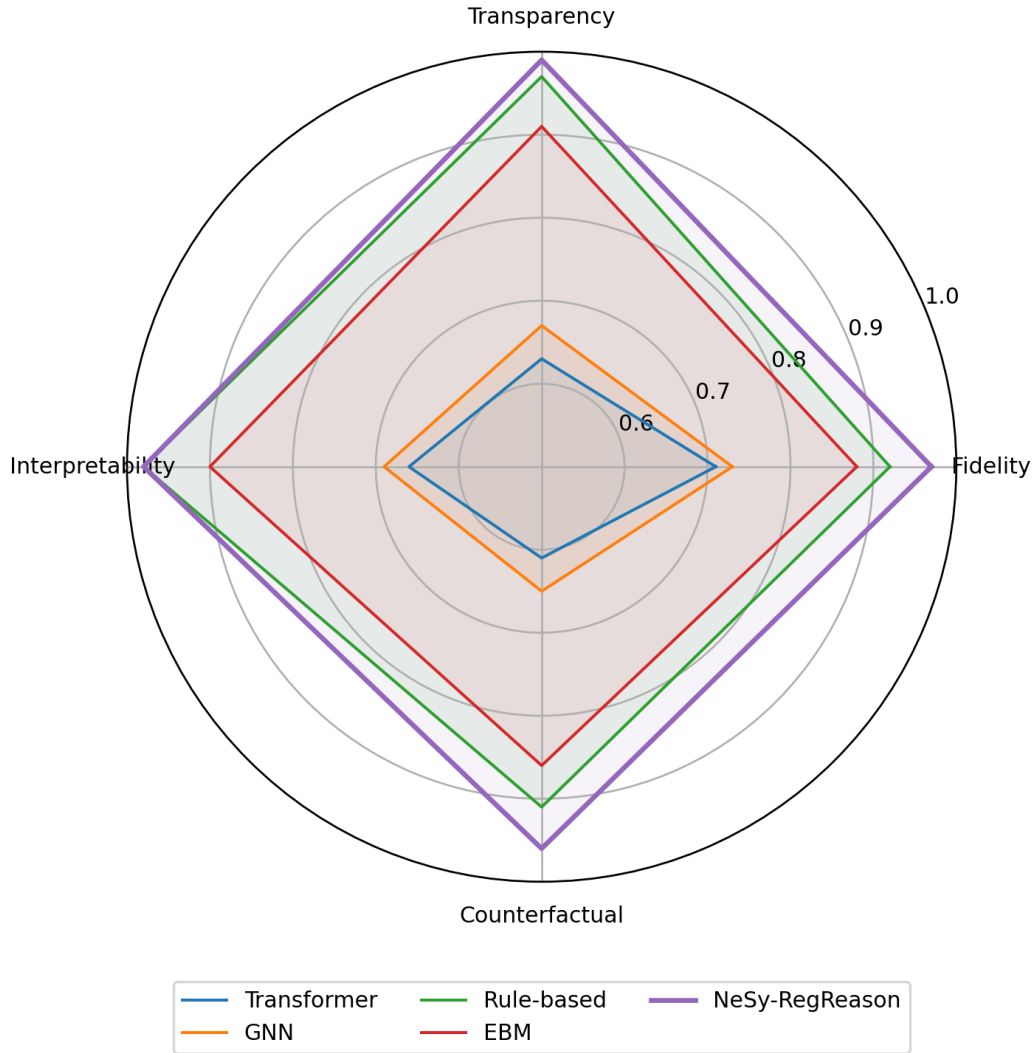


Figure 3. Explainability profile of neural, symbolic, interpretable, and neuro-symbolic models.

NeSy-RegReason reached 0.97 fidelity, 0.99 transparency, 0.98 interpretability, and 0.96 counterfactual quality. The rule-based system matched interpretability at 0.98 and approached transparency, confirming that symbolic structure is naturally inspectable. The proposed framework added higher fidelity and counterfactual quality while retaining neural adaptability. Compared with the Transformer, the absolute improvements ranged from 0.26 for

fidelity to 0.36 for transparency. These results support H2 and show why explanation should be measured as a profile rather than a single score.

Table 6. Auditability and regulatory conflict resolution

Method	Trace complete %	Reasoning consistent %	Reproducible %	Evidence quality %	Conflict resolution %
Rule-Based System	93.52	94.41	93.84	92.61	91.32
Transformer	68.24	71.32	70.85	66.42	73.84
GNN	71.83	73.15	72.68	69.37	76.54
EBM	88.71	89.82	88.46	87.54	88.15
NeSy-RegReason	99.12	98.64	98.81	98.32	98.54

Audit-trace completeness reached 99.12%, reasoning consistency 98.64%, decision reproducibility 98.81%, and evidence quality 98.32%. Conflict resolution reached 98.54%, with 99.12% rule consistency and 98.71% legal-reasoning accuracy in the supplied detailed results. These values support H5 and H6. The remaining gap to 100% is operationally important. An incomplete trace should trigger an exception rather than be treated as acceptable merely because aggregate completeness is high.

4.3 Adaptability, ablation, and significance

Table 7. Adaptability to regulatory change

Method	Rule update (s)	Recovery (s)	Retraining cost	Update success %
Random Forest	352	498	High	81.4
XGBoost	301	437	High	84.2
Transformer	278	396	High	87.1
GNN	261	382	High	88.3
Rule-Based System	188	241	Medium	93.2
NeSy-RegReason	96	124	Low	98.6

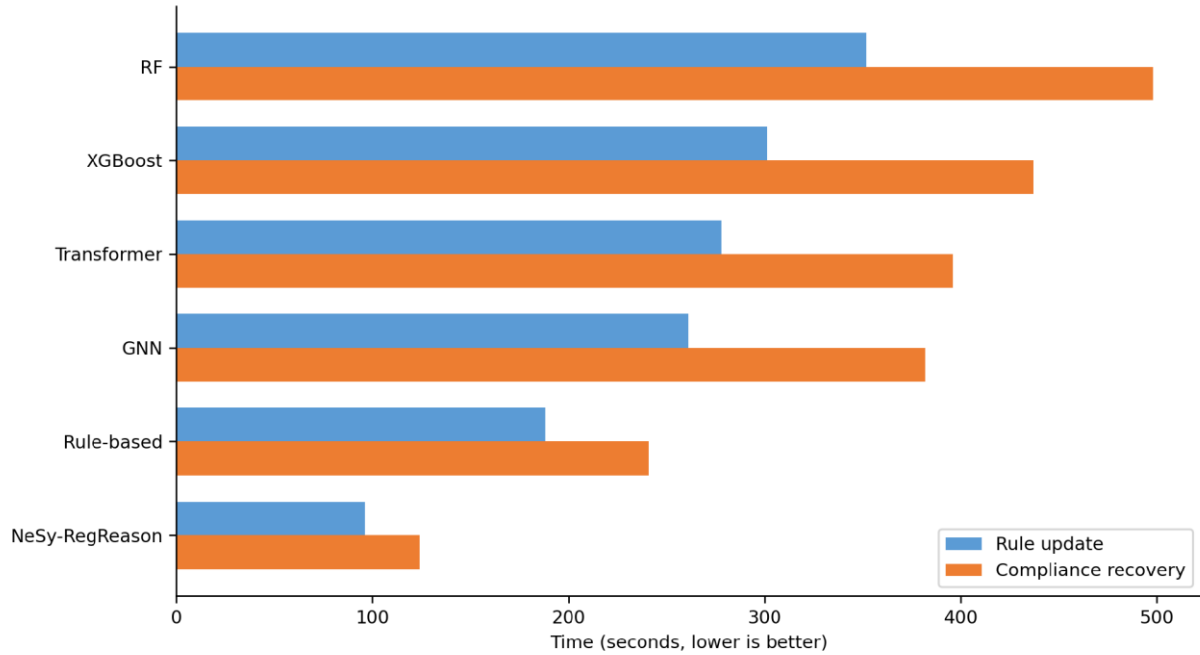


Figure 4. Rule-update and compliance-recovery latency under regulatory change.

NeSy-RegReason updated rules in 96 seconds and recovered compliance operation in 124 seconds. Relative to the rule-based system, update time declined by 49% and recovery time by 49%. Relative to the GNN, the reductions were 63% and 68%, respectively. The result follows from separating rule changes from full model retraining: many normative changes can be represented in the knowledge and rule layer, while the neural component is fine-tuned only when feature semantics or behavioural distributions change. Knowledge-update success of 98.6% supports H4, subject to disclosure of the change scenarios and acceptance criteria.

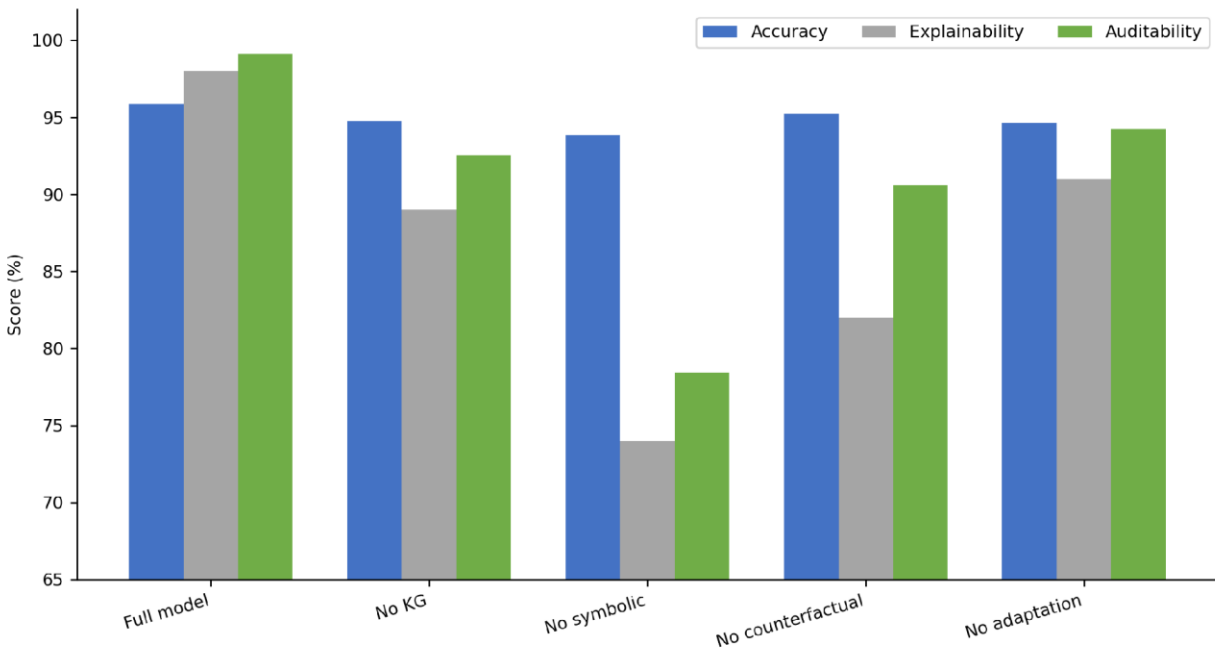


Figure 5. Ablation analysis of accuracy, explainability, and auditability.

Removing symbolic reasoning caused the largest deterioration: accuracy fell to 93.82%, explainability to 0.74, and auditability to 78.43%. Removing the knowledge graph reduced explanation and audit scores to 0.89 and 92.54%; removing counterfactual generation primarily affected explanation quality; and removing adaptation affected accuracy under changed rules. The pattern indicates complementarity rather than a single dominant predictive module. The Wilcoxon tests reported $p < .001$ against XGBoost, Transformer, GNN, and the rule-based system, and $p = .002$ against EBM. Effect sizes ranged from 0.68 to 0.83. These values are consistent with meaningful differences, but the absence of paired observations prevents independent verification or assessment of multiplicity.

5. Discussion

5.1 Interpretation and contribution

The central finding is that regulatory compliance benefits from changing the role of machine learning. In a conventional classifier, the predicted label is the decision and an explanation is generated afterwards. In NeSy-RegReason, the predicted score is evidence submitted to a legal-reasoning process. This distinction explains the joint improvement in detection, false alerts, rule satisfaction, explanation fidelity, and auditability. Neural models contribute sensitivity to complex and evolving patterns; symbolic models contribute normative structure, exceptions, consistency checks, and provenance. The knowledge graph supplies the shared interface between them.

The 95.86% accuracy result alone would be an incremental gain over strong neural baselines. The more substantial contribution is the combination of 98.31% detection, 1.89% false positives, 0.97 explanation fidelity, 99.12% trace completeness, and 96-second rule updates. These outcomes address different institutional stakeholders. Analysts need fewer low-value alerts; compliance officers need applicable rules and exceptions; internal audit needs reproducibility; model-risk teams need versions and calibration; regulators need evidence paths; and engineering teams need bounded adaptation latency. A single accuracy metric cannot represent these needs.

The conflict module is particularly important in cross-border finance. A transaction may simultaneously invoke AML monitoring, privacy restrictions, consumer-protection duties, sanctions screening, and local reporting timelines. Simple rule conjunction can produce contradictory actions. Encoding authority, jurisdiction, specificity, effective date, and exception relations allows the system to state not only what conclusion it reached but why one obligation prevailed. This is closer to legal operationalisation than generic feature attribution.

5.2 Practical deployment model

Deployment should use a controlled three-lane decision policy. Low-risk cases with complete evidence and no conflicts may be automatically cleared. High-risk cases with satisfied mandatory rules and complete traces may be escalated with a machine-generated case package. Cases with missing evidence, unresolved conflict, low calibration confidence, novel graph structure, or failed trace validation must abstain and enter human review. Human decisions should be captured as labelled feedback, but they should not automatically modify legal rules. Rule changes require governance approval, test cases, effective dates, rollback capability, and independent validation.

A production audit record should include: transaction and entity snapshot identifiers; feature-schema version; neural model and calibration version; knowledge-graph and ontology version; every activated rule with source clause and effective date; grounded facts; inference and conflict-resolution steps; confidence components; counterfactual; final outcome; reviewer action; and cryptographic integrity metadata. Retention and access must follow privacy and sector-specific requirements. GDPR data minimisation means that an audit-ready trace should store necessary evidence and stable references, not indiscriminately duplicate personal data [3].

5.3 Governance and responsible AI implications

Neuro-symbolic architecture does not eliminate model risk. Rules may be incorrectly encoded, ontologies may omit legally relevant concepts, neural models may be biased, and apparently complete traces may faithfully reproduce a flawed decision. Governance therefore requires separate validation of statistical performance, legal correctness, data

representativeness, calibration, security, and human factors. The NIST AI Risk Management Framework and ISO/IEC 42001 provide useful governance structures for mapping, measuring, managing, and monitoring AI risk [33], [34]. Model documentation should also distinguish intended use, prohibited use, known limitations, and performance across customer and jurisdictional subgroups [35].

Counterfactual explanations need special care. A technically minimal change may be legally impossible, ethically inappropriate, or operationally misleading. For example, suggesting a different country, age, or customer identity is not an actionable compliance explanation. The counterfactual generator must therefore constrain immutable and protected attributes, preserve causal plausibility, and identify whether the change concerns customer behaviour, missing documentation, institutional procedure, or regulatory scope. Human-readable language should be generated from verified rule templates to avoid adding unsupported legal claims.

5.4 Limitations

The study has six material limitations. First, aggregate results were author-supplied and could not be independently recomputed. Second, the synthetic compliance dataset lacks a published generator and expert-validation protocol. Third, public fraud datasets are proxies for regulatory compliance and may not reflect operational base rates, investigation workflows, or legal labels. Fourth, explainability and audit metrics require validated instruments and multiple expert raters; inter-rater reliability was not supplied. Fifth, adaptation times depend on hardware, rule complexity, graph size, and change scenarios, which were not documented. Sixth, the framework spans regulations with different legal purposes, so high aggregate coverage can conceal weak performance for a particular jurisdiction or obligation.

Consequently, the results should be interpreted as a strong proof-of-concept dataset rather than final evidence of production readiness. A submission should include a repository containing code, ontology, rules, synthetic-data generator, splits, seeds, hyperparameters, raw predictions, statistical notebook, model cards, and anonymised expert-rating forms. External legal review and temporal validation on later regulatory versions are also necessary.

5.5 Robustness, security, and economic evaluation

A compliance engine is exposed to strategic behaviour. Adversaries may split transactions, rotate counterparties, manipulate narrative fields, poison feedback, or exploit differences between jurisdictions. Robustness testing should therefore include transaction structuring, synthetic identity linkage errors, graph edge deletion, timestamp manipulation, missing-document attacks, and adversarial text. The neural layer should be evaluated under distribution shift, while the symbolic layer should be tested against inconsistent facts, cyclic rules, missing priorities, and deliberately malformed updates. A secure system must fail closed for missing mandatory evidence without converting every technical error into an accusation against a customer.

Knowledge and rule infrastructure also expands the attack surface. An unauthorised rule edit could suppress alerts or generate mass false positives while leaving the predictive model unchanged. Production controls should include least-privilege access, multi-person approval, signed rule bundles, integrity verification, tamper-evident logs, dependency scanning, secret isolation, and continuous reconciliation between approved and deployed versions. Red-team exercises should test whether a user can prompt an explanation component to disclose restricted data or to generate language inconsistent with the verified inference. Natural-language generation should never be allowed to modify the underlying proof.

Operational benefit should be evaluated with cost-sensitive measures. False positives consume analyst capacity and may delay legitimate transactions; false negatives create enforcement, loss, and reputational exposure; abstentions create review work but can prevent unjustified automation. A realistic study should estimate expected cost using institution-specific weights, measure alerts per thousand transactions, median investigation time, escalation rate, reviewer disagreement, and percentage of cases closed without additional information. The 1.89% false-positive rate is promising, but its economic meaning depends on prevalence, transaction volume, staffing, and the severity distribution of missed cases.

Scalability should be measured separately for neural scoring, graph retrieval, rule execution, explanation construction, and audit persistence. Mean latency is insufficient because compliance services face bursty payment traffic; p95 and p99 latency, throughput, memory, graph growth, and recovery under component failure are necessary. Rule-update time should be decomposed into parsing, validation, reasoning consistency checks, test execution, approval, and deployment. Such decomposition would clarify whether the reported 96 seconds represents machine processing alone or the complete governed change process.

Finally, subgroup and jurisdiction analysis is necessary even where protected attributes are unavailable. Error rates can be compared across customer risk tiers, products, channels, geographies, transaction values, graph connectivity, new versus established customers, and data-completeness bands. Fairness conclusions require careful legal and causal interpretation because some risk factors may be legitimate for one obligation and prohibited for another. NeSy-RegReason can support this analysis by recording which rules and evidence drove each outcome, but transparency is an enabling condition rather than proof of fairness.

6. Conclusion and Future Work

NeSy-RegReason integrates neural prediction with a regulatory knowledge graph, symbolic inference, conflict resolution, constrained counterfactual explanation, adaptive rule updating, and audit-trace generation. On the supplied multi-dataset results, it outperformed traditional machine learning, neural, symbolic, and interpretable baselines, reaching 95.86% accuracy, 0.986 ROC-AUC, 98.31% compliance detection, 1.89% false positives, 0.97 explanation fidelity, 99.12% audit completeness, and 98.54% conflict resolution. Ablations show that symbolic reasoning is indispensable for explainability and auditability, while the knowledge graph, counterfactual module, and adaptation mechanism make distinct contributions.

Future work should conduct preregistered, multi-institution evaluation with jurisdiction-specific test suites and temporal holdouts; release the ontology, rules, code, and synthetic generator; measure calibration, fairness, robustness, energy use, and analyst workload; compare alternative non-monotonic and probabilistic logics; and test human-regulator comprehension through controlled studies. Further research should also examine secure rule updates, privacy-preserving graph learning, formal verification of conflict policies, multilingual regulation extraction, and continuous monitoring for both behavioural and legal drift. The framework’s practical value will ultimately depend not on high aggregate scores alone, but on whether every consequential decision can be reproduced, legally justified, challenged, and safely updated.

REFERENCES

1. Financial Action Task Force, International Standards on Combating Money Laundering and the Financing of Terrorism & Proliferation: The FATF Recommendations, updated Feb. 2025.
2. Basel Committee on Banking Supervision, Principles for the Sound Management of Operational Risk, Bank for International Settlements, 2021.
3. European Parliament and Council, Regulation (EU) 2016/679 (General Data Protection Regulation), 2016.
4. European Parliament and Council, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence, Official Journal of the European Union, 2024.
5. T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” Proc. KDD, pp. 785–794, 2016. doi:10.1145/2939672.2939785.
6. G. Ke et al., “LightGBM: A highly efficient gradient boosting decision tree,” Advances in Neural Information Processing Systems, vol. 30, 2017.
7. L. Prokhorenkova et al., “CatBoost: Unbiased boosting with categorical features,” Advances in Neural Information Processing Systems, vol. 31, 2018.
8. A. Vaswani et al., “Attention is all you need,” Advances in Neural Information Processing Systems, vol. 30, 2017.
9. Z. Wu et al., “A comprehensive survey on graph neural networks,” IEEE Trans. Neural Networks Learn. Syst., vol. 32, no. 1, pp. 4–24, 2021. doi:10.1109/TNNLS.2020.2978386.
10. Financial Stability Institute, “Regulating AI in the financial sector: recent developments and main challenges,” FSI Insights, Bank for International Settlements, 2025.
11. W3C, OWL 2 Web Ontology Language Document Overview, 2nd ed., W3C Recommendation, 2012.
12. I. Horrocks et al., “SWRL: A semantic web rule language combining OWL and RuleML,” W3C Member Submission, 2004.

13. A. Colmerauer and P. Roussel, "The birth of Prolog," in *History of Programming Languages II*, ACM, 1996, pp. 331–367.
14. E. Sirin et al., "Pellet: A practical OWL-DL reasoner," *Journal of Web Semantics*, vol. 5, no. 2, pp. 51–53, 2007.
15. A. Hogan et al., "Knowledge graphs," *ACM Computing Surveys*, vol. 54, no. 4, art. 71, 2022. doi:10.1145/3447772.
16. A. d'Avila Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave," *Artificial Intelligence Review*, vol. 56, pp. 12387–12406, 2023.
17. M. Raedt et al., "Neuro-symbolic artificial intelligence: The state of the art," IOS Press, 2022.
18. U. Nawaz et al., "A review of neuro-symbolic AI integrating reasoning and learning," *Machine Learning with Applications*, 2025.
19. K. Hamilton et al., "Neuro-symbolic artificial intelligence: a survey," *Frontiers in Artificial Intelligence*, vol. 7, 2024.
20. K. Khanvilkar and K. Kommuru, "Regulatory graphs and GenAI for real-time transaction monitoring and compliance explanation in banking," arXiv:2506.01093, 2025.
21. Y. S. Hsia, F. Yu, and J.-H. R. Jiang, "Neuro-symbolic compliance: Integrating LLMs and SMT for automated financial legal analysis," *AIware*, 2025.
22. A. Akinfaderin and S. Subramanian, "VERAFI: Verified agentic financial intelligence through neurosymbolic policy generation," arXiv:2512.14744, 2025.
23. Stanford CodeX, "Neuro-Symbolic AI for Regulatory Compliance: The ComplianceTwin Pilot," Stanford Law School, 2025.
24. M. Weber et al., "Anti-money laundering in Bitcoin: Experimenting with graph convolutional networks for financial forensics," *KDD Workshop on Anomaly Detection in Finance*, 2019.
25. E. A. Lopez-Rojas, A. Elmir, and S. Axelsson, "PaySim: A financial mobile money simulator for fraud detection," *European Modeling and Simulation Symposium*, 2016.
26. IEEE Computational Intelligence Society, "IEEE-CIS Fraud Detection," Kaggle competition dataset, 2019.
27. M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?' Explaining the predictions of any classifier," *Proc. KDD*, pp. 1135–1144, 2016.
28. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
29. H. Nori et al., "InterpretML: A unified framework for machine learning interpretability," arXiv:1909.09223, 2019.
30. S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box," *Harvard Journal of Law & Technology*, vol. 31, no. 2, pp. 841–887, 2018.
31. J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 7, pp. 1–30, 2006.
32. S. García and F. Herrera, "An extension on statistical comparisons of classifiers over multiple data sets," *Journal of Machine Learning Research*, vol. 9, pp. 2677–2694, 2008.
33. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023.
34. ISO/IEC, *ISO/IEC 42001:2023 Information technology—Artificial intelligence—Management system*, 2023.
35. M. Mitchell et al., "Model cards for model reporting," *Proc. FAT**, pp. 220–229, 2019. doi:10.1145/3287560.3287596.
36. European Banking Authority, *Report on Machine Learning for IRB Models*, EBA/REP/2021/09, 2021.
37. European Commission High-Level Expert Group on AI, *Assessment List for Trustworthy Artificial Intelligence (ALTAI)*, 2020.
38. N. Mehrabi et al., "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, art. 115, 2021.