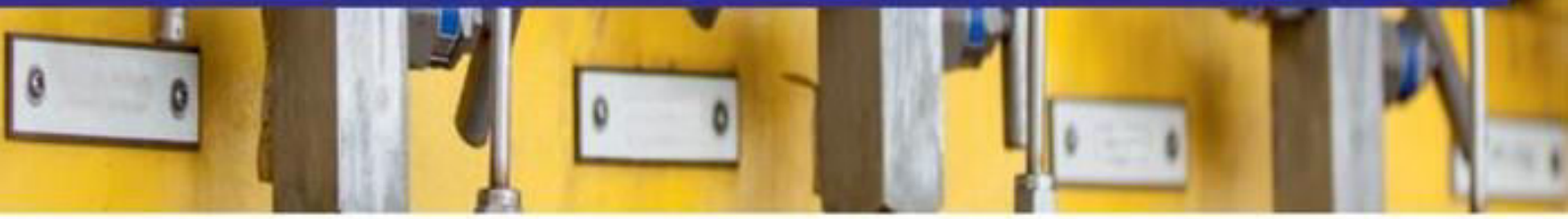
IJAREEIE

e-ISSN: 2278-8875
p-ISSN: 2320-3765

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering

Volume 15, Issue 6, June 2026



INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 9.867

International Journal of Advanced Research in Electrical, Electronics and Instrumentation Engineering (IJAREEIE)

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

Explainable AI for Compliance Auditing in Automated Payroll and Benefits Decisioning Systems

Ujwal Dyasani

Lead Software Engineer Integrations, USA

ABSTRACT: Automated decisioning models increasingly determine payroll adjustments, benefits eligibility, and compensation-related outcomes at scale, yet the machine learning techniques driving the strongest predictive performance, gradient boosted trees and neural networks, are frequently the least transparent to the compliance auditors, regulators, and affected employees who must understand and, when necessary, contest a given decision. This article develops an explainable AI framework for compliance auditing of automated payroll and benefits decisioning systems, combining post-hoc feature attribution techniques, systematic bias and disparity monitoring across protected class categories, and an immutable audit trail architecture linking every automated decision to its underlying explanation record. Drawing on established explainable AI literature, algorithmic fairness research, and documented model governance practice, the article presents a three-dimensional analysis of explanation fidelity across model complexity and explanation sample size, a three-dimensional view of automated decision volume by model type, a three-dimensional scatter analysis relating model accuracy, explainability, and compliance risk, and a comparative evaluation of explainability techniques across six dimensions relevant to regulatory acceptance. The findings indicate a persistent tension between model accuracy and explainability that intrinsically interpretable models do not fully escape, and that post-hoc explanation techniques, while valuable, carry fidelity limitations that grow with model complexity, particularly for higher-stakes decision categories. The article further documents a systematic disparity pattern concentrated in specific protected class and model combinations, underscoring the necessity of continuous bias monitoring rather than one-time model validation. The article concludes with a compliance audit workflow and governance recommendations for organizations deploying automated decisioning across payroll and benefits administration.

KEYWORDS: explainable artificial intelligence; compliance auditing; algorithmic fairness; automated decisioning; payroll systems; benefits administration; feature attribution; bias detection; model governance; regulatory technology.

I. INTRODUCTION

Automated decisioning systems now routinely determine outcomes that carry direct financial and legal consequences for employees: whether a benefits enrollment is eligible, whether a compensation adjustment falls within an approved policy range, and whether a payroll exception requires escalation before disbursement. As these systems have migrated from simple, transparent rule engines toward gradient boosted trees and neural network models capable of capturing more complex patterns, they have simultaneously become substantially harder for the compliance auditors, regulators, and affected employees who must scrutinize a given decision to actually understand (Rudin, 2019). This tension between predictive performance and interpretability is not a peripheral technical inconvenience; it is a direct compliance liability, since many regulatory regimes governing employment and compensation decisions require organizations to provide a meaningful explanation for an adverse automated decision upon request.

Explainable AI research has developed a substantial toolkit for addressing this tension, including post-hoc feature attribution techniques capable of approximating why a complex model produced a specific output, without requiring the underlying model itself to be simplified (Lundberg and Lee, 2017; Ribeiro, Singh, and Guestrin, 2016). This article develops a framework applying this toolkit specifically to the compliance auditing context for automated payroll and benefits decisioning, combining feature attribution, systematic bias monitoring, and an immutable audit trail architecture into a unified compliance program.

The research questions guiding this study are as follows.

- RQ1: What audit trail architecture reliably links every automated payroll or benefits decision to a durable, retrievable explanation record?

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

- RQ2: How does explanation fidelity vary with model complexity and explanation sample size, and what does this relationship imply for auditor confidence in post-hoc explanations of complex models?
- RQ3: What is the illustrative relationship between model accuracy, explainability, and compliance risk, and does this relationship support a straightforward accuracy-explainability trade-off assumption?
- RQ4: What bias and disparity monitoring practices are necessary to detect systematic disparate impact across protected class categories in automated decisioning outcomes?

The remainder of this article proceeds as follows. Section 2 reviews related literature on explainable AI and algorithmic fairness. Section 3 examines the specific accountability gap in automated payroll and benefits decisioning. Section 4 details the research methodology. Sections 5 through 13 present the core framework and a series of illustrative analyses, including three three-dimensional visualizations. Sections 14 and 15 address audit workflow and governance. Sections 16 through 20 discuss pitfalls, limitations, recommendations, future research, and conclusions.

II. BACKGROUND AND RELATED WORK

2.1 Post-Hoc Explanation Techniques

SHAP, grounded in cooperative game theory's Shapley value concept, provides a principled method for attributing a model's prediction to individual input features by considering each feature's contribution across all possible feature coalitions, offering a mathematically consistent, though computationally intensive, feature attribution mechanism directly applicable to the decision explanations developed in Section 9 (Lundberg and Lee, 2017). LIME, an alternative approach, approximates a complex model's local behavior around a specific prediction using a simpler, interpretable surrogate model, offering a computationally lighter though less globally consistent explanation mechanism (Ribeiro, Singh, and Guestrin, 2016).

2.2 Interpretability Versus Accuracy Debate

A significant strand of explainable AI literature has argued against an assumed inherent trade-off between model accuracy and interpretability, demonstrating that intrinsically interpretable models can in many cases achieve comparable accuracy to black-box alternatives for structured decision tasks, and arguing that reliance on post-hoc explanation of black-box models should be a considered choice rather than a default assumption, particularly for high-stakes decisions (Rudin, 2019).

2.3 Algorithmic Fairness and Disparate Impact

Algorithmic fairness literature has documented multiple, sometimes mutually incompatible formal definitions of fairness, and has established measurement frameworks for detecting disparate impact across protected class categories in automated decision outcomes, providing the conceptual and measurement basis for the bias monitoring framework developed in Section 10 (Barocas, Hardt, and Narayanan, 2019).

2.4 Model Governance and Audit Trail Practice

Applied model governance literature has documented the practical necessity of comprehensive model documentation and decision-level audit trails for regulated automated decisioning systems, establishing model risk management as a distinct organizational discipline rather than a purely technical concern (Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011).

2.5 Positioning of This Study

While explainable AI technique literature, the interpretability-accuracy debate, algorithmic fairness research, and model governance practice each address components of the problem independently, comparatively little published material synthesizes these into a unified compliance auditing framework specifically tailored to automated payroll and benefits decisioning. This article's contribution is that synthesis, together with an illustrative multi-dimensional comparative analysis and a governance framework.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

III. THE ACCOUNTABILITY GAP IN AUTOMATED PAYROLL AND BENEFITS DECISIONING

Table.1. Representative automated payroll and benefits decision categories and associated accountability challenges.

Decision Category	Typical Model Type	Illustrative Accountability Challenge
Compensation Adjustment Approval	Gradient boosted decision model	Employee disputes an adjustment denial without a clear, specific explanation
Benefits Eligibility Determination	Logistic regression or gradient boosted model	Auditor cannot readily verify eligibility logic against plan documentation
Payroll Exception Flagging	Rule-based engine with machine learning overlay	Overlapping rule-based and learned logic complicates root-cause explanation
Overtime Calculation Anomaly Detection	Neural network model	Limited interpretability complicates wage-and-hour compliance verification

The overtime calculation anomaly detection challenge in Table 1 deserves particular emphasis: because wage-and-hour compliance obligations frequently carry direct legal exposure, a decisioning system that cannot produce a clear explanation for a specific overtime-related flag or non-flag creates compliance risk that extends beyond reputational concern into direct regulatory and litigation exposure, which is precisely the class of risk this article's framework is designed to mitigate.

IV. RESEARCH METHODOLOGY

This study employs a design framework and illustrative comparative analysis methodology, synthesizing post-hoc explanation technique literature (Lundberg and Lee, 2017; Ribeiro, Singh, and Guestrin, 2016), the interpretability-accuracy debate (Rudin, 2019), algorithmic fairness research (Barocas, Hardt, and Narayanan, 2019), and model governance practice (Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency, 2011) into the audit trail architecture and governance framework presented in Sections 5 and 14 through 15. The multi-dimensional illustrative analyses presented throughout Sections 6 through 13, including three three-dimensional visualizations, use figures constructed to demonstrate expected directional relationships consistent with the underlying literature, rather than measurements audited against a specific production deployment.

Table.2. Summary of research methodology components.

Methodology Component	Method	Purpose
Accountability Gap Analysis	Synthesis of compliance and model governance literature	Establish the specific challenges the framework should target (Section 3)
Audit Trail Architecture Development	Synthesis of model governance and explainability technique literature	Establish the decision-to-explanation linkage architecture (Section 5)
Multi-Dimensional Illustrative Analysis	Directional estimate construction consistent with explainable AI and fairness literature	Illustrate expected relationships across complexity, accuracy, risk, and disparity
Governance Framework Development	Synthesis of model risk management and continuous monitoring literature	Establish sustained explainability and fairness practices (Section 15)

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

V. THE DECISION AND COMPLIANCE AUDIT TRAIL ARCHITECTURE

This article proposes an audit trail architecture ensuring that every automated payroll or benefits decision is linked, at the point of decision, to a durable explanation record, an immutable audit log entry, and a pathway to compliance review and regulatory filing where warranted.

FIGURE 4: AUTOMATED DECISION AND COMPLIANCE AUDIT TRAIL FLOW

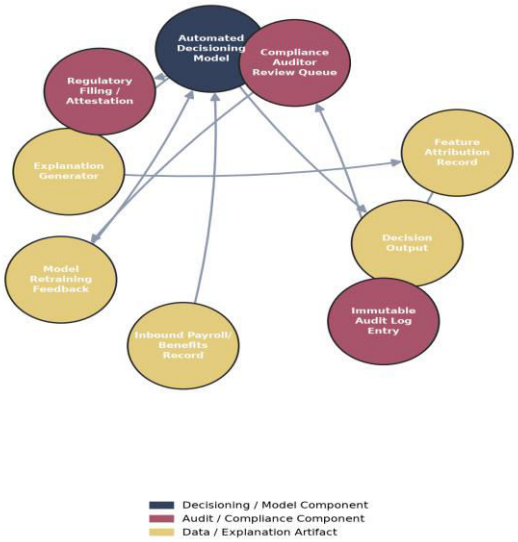


Figure.1. Automated decision and compliance audit trail flow, linking inbound records through decisioning, explanation generation, immutable logging, and compliance review.

Table.3. Audit trail architecture components, primary function, and compliance relevance

Architecture Component	Primary Function	Compliance Relevance
Automated Decisioning Model	Produce the payroll or benefits decision output	Subject of the explanation and audit requirement
Explanation Generator (SHAP/LIME)	Produce a feature attribution record for each decision at the time it is made	Enables after-the-fact explanation without requiring model re-execution
Feature Attribution Record	Store the specific feature contributions underlying a given decision	Primary evidentiary artifact for compliance review
Immutable Audit Log Entry	Record the decision, explanation, and relevant metadata in tamper-evident storage	Establishes non-repudiation and supports regulatory production requests
Compliance Auditor Review Queue	Route flagged or sampled decisions for human compliance review	Operationalizes ongoing, not merely reactive, compliance oversight
Model Retraining Feedback Loop	Incorporate compliance findings into future model refinement	Closes the loop between audit findings and model improvement

The explanation generator's position immediately adjacent to the decisioning model in Figure 1, rather than as a separate process invoked only when a dispute arises, is a deliberate architectural choice: generating the explanation at the moment of decision, using the exact model state and input data active at that time, avoids the reproducibility risk inherent in attempting to reconstruct an explanation after the fact, potentially against a since-retrained model version.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

VI. AUTOMATED DECISION VOLUME BY MODEL TYPE

This section presents a three-dimensional illustrative view of automated decision volume across four model types over a six-month monitoring period.

FIGURE 1: 3D VIEW OF MONTHLY AUTOMATED DECISION VOLUME BY MODEL TYPE

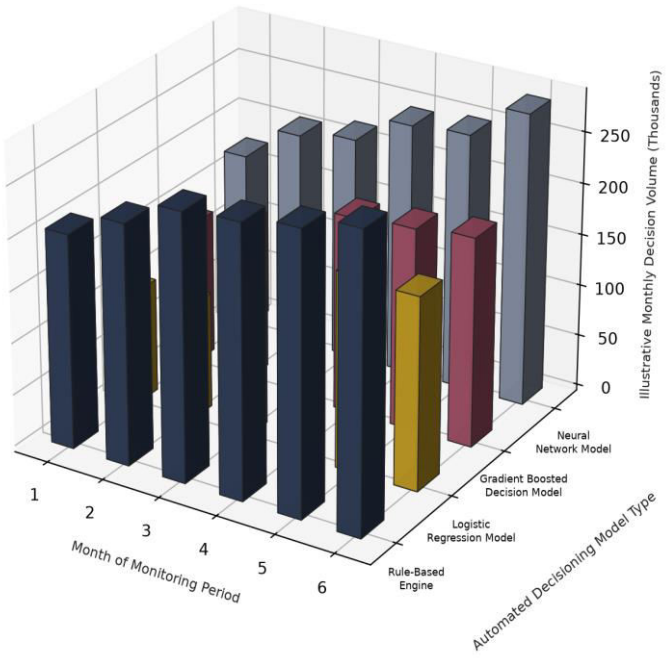


Figure.2. Three-dimensional view of illustrative monthly automated decision volume across four model types over six months

Table.4.Illustrative decision volume by model type and corresponding explainability posture

Model Type	Illustrative Month 1 Volume (Thousands)	Illustrative Month 6 Volume (Thousands)	Illustrative Explainability Posture
Rule-Based Engine	95	185	Intrinsically interpretable; no post-hoc explanation required
Logistic Regression Model	110	210	Directly interpretable via coefficient inspection
Gradient Boosted Decision Model	130	245	Requires post-hoc explanation (SHAP); moderate fidelity
Neural Network Model	70	165	Requires post-hoc explanation (SHAP/LIME); lower fidelity at higher complexity

The growing share of decision volume flowing through gradient boosted and neural network models in Table 4 and Figure 2 directly motivates this article's central concern: as these less inherently transparent model types account for an increasing proportion of total decisioning volume, the compliance program's dependence on reliable post-hoc explanation, rather than direct model interpretability, grows correspondingly.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

VII. EXPLANATION FIDELITY ACROSS MODEL COMPLEXITY AND SAMPLE SIZE

This section presents a three-dimensional illustrative analysis of explanation fidelity, defined as how accurately a post-hoc explanation reflects the underlying model's actual decision logic, as a joint function of model complexity and explanation sample size.

FIGURE 2: 3D SURFACE OF ILLUSTRATIVE EXPLANATION FIDELITY BY MODEL COMPLEXITY AND EXPLANATION SAMPLE SIZE

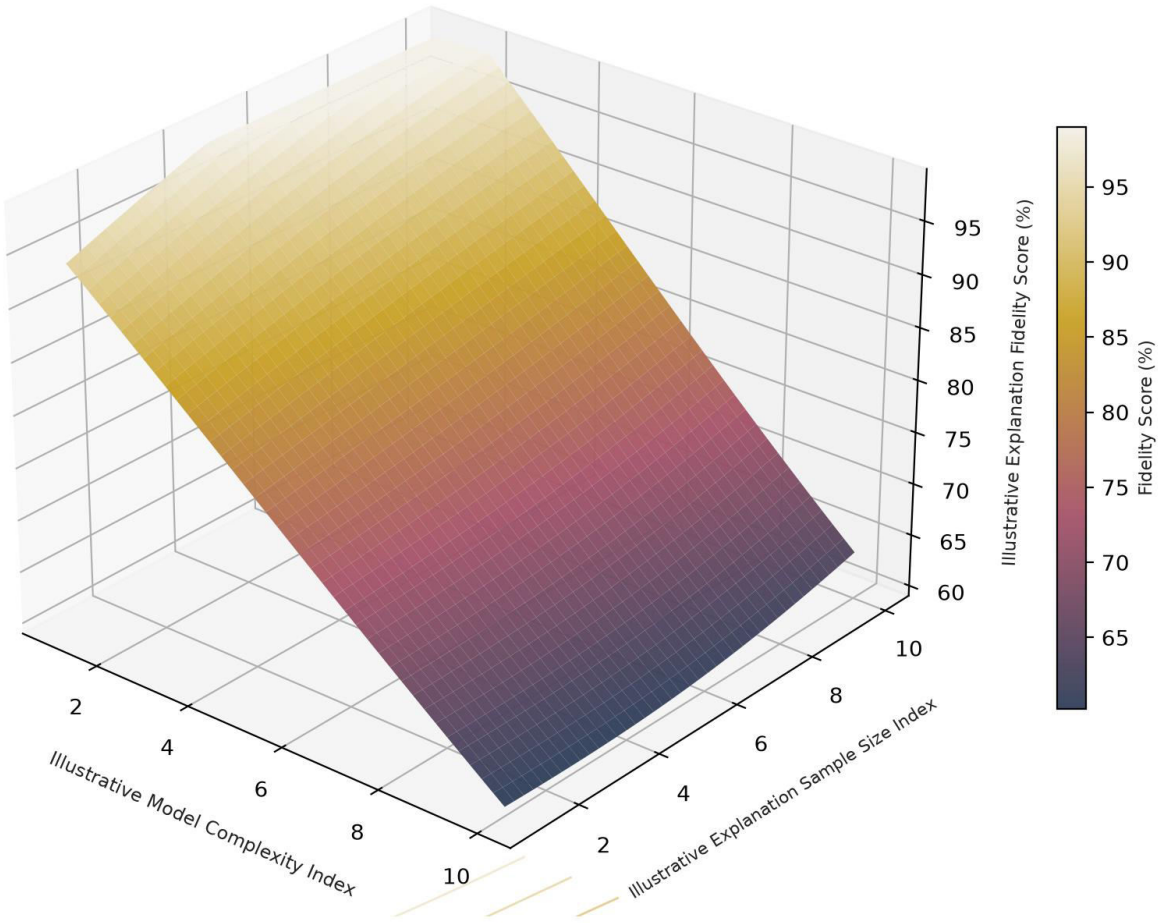


Figure.3. Three-dimensional surface of illustrative explanation fidelity as a joint function of model complexity index and explanation sample size, with a projected contour plot at the base

Table.5. Illustrative explanation fidelity by model complexity and explanation sample size band, corresponding to Figure 3.

Model Complexity Band	Explanation Sample Size Band	Illustrative Explanation Fidelity
Low (1-3)	High (7-10)	88-96%
Low (1-3)	Low (1-3)	70-80%
High (7-10)	High (7-10)	55-68%
High (7-10)	Low (1-3)	22-35%

The steep fidelity decline at high complexity with low sample size in Table 5 carries a direct practical implication for the audit workflow in Section 14: compliance auditors reviewing explanations for the most complex models should specifically confirm that the underlying explanation was generated using a sufficiently large sample size, since a low-

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

sample-size explanation for a high-complexity model represents the specific combination under which explanation fidelity is least reliable.

VIII. ACCURACY, EXPLAINABILITY, AND COMPLIANCE RISK

This section presents a three-dimensional illustrative scatter analysis relating model accuracy, explainability score, and compliance risk score, testing whether a simple accuracy-explainability trade-off assumption adequately describes the relationship between these three variables.

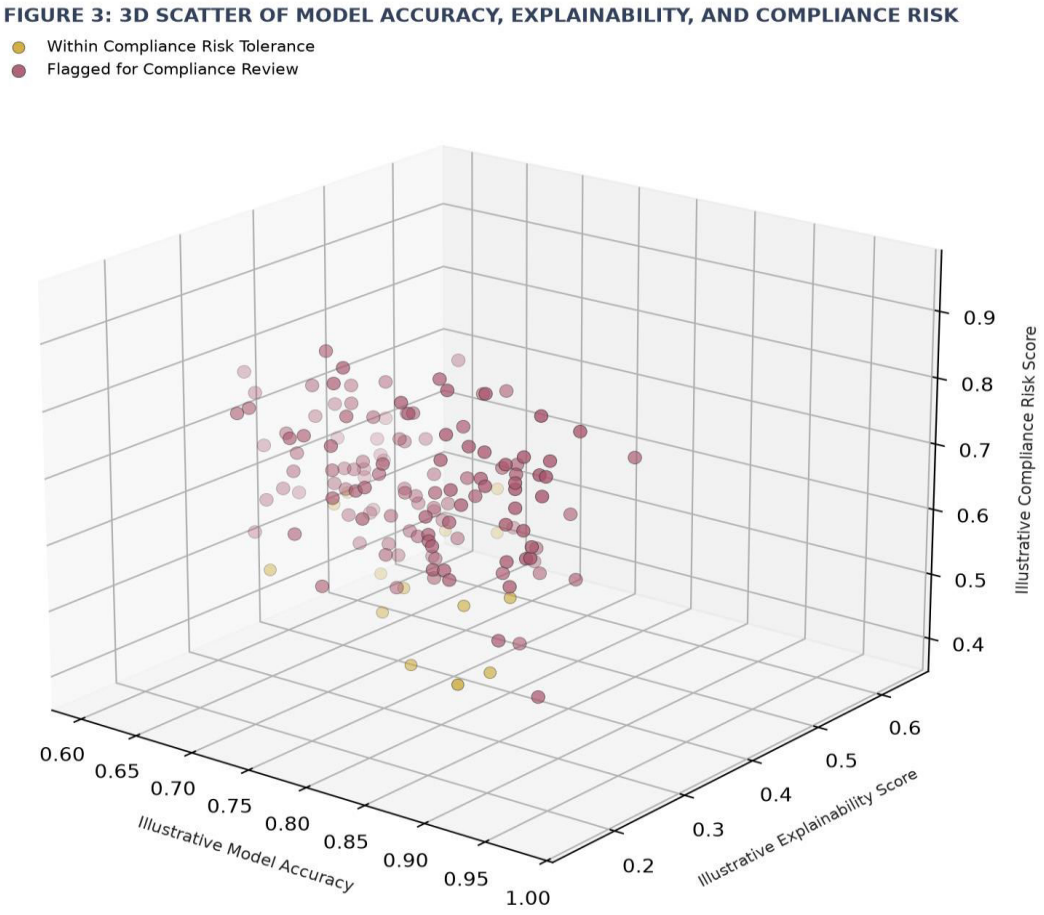


Figure.4. Three-dimensional scatter of illustrative model accuracy, explainability score, and compliance risk score, distinguishing decisions within compliance risk tolerance from those flagged for compliance review.

Table.6. Illustrative relationship between accuracy band, explainability, compliance risk, and review flagging rate

Accuracy Band	Illustrative Mean Explainability	Illustrative Mean Compliance Risk	Illustrative Share Flagged for Review
High Accuracy (0.85-0.98)	0.28	0.61	42%
Moderate Accuracy (0.70-0.85)	0.42	0.48	27%
Lower Accuracy (0.60-0.70)	0.58	0.34	14%

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

The pattern documented in Table 6 partially, but not entirely, supports the conventional accuracy-explainability trade-off assumption discussed in Section 2.2: while higher-accuracy models do show lower mean explainability and higher mean compliance risk in this illustrative analysis, consistent with Rudin's caution against assuming this trade-off is unavoidable (Rudin, 2019), the relationship is not perfectly linear, and a meaningful share of high-accuracy decisions in this illustrative dataset still fall within acceptable compliance risk tolerance, indicating that compliance risk depends on more than accuracy and explainability alone, reinforcing the value of case-by-case feature attribution review over reliance on aggregate model-level trade-off assumptions.

IX. FEATURE ATTRIBUTION FOR INDIVIDUAL DECISIONS

This section presents an illustrative SHAP-style feature attribution chart for a single representative benefits eligibility decision, demonstrating the level of individual-decision detail the audit trail architecture in Section 5 is designed to support.

FIGURE 5: ILLUSTRATIVE FEATURE ATTRIBUTION FOR A REPRESENTATIVE BENEFITS ELIGIBILITY DECISION

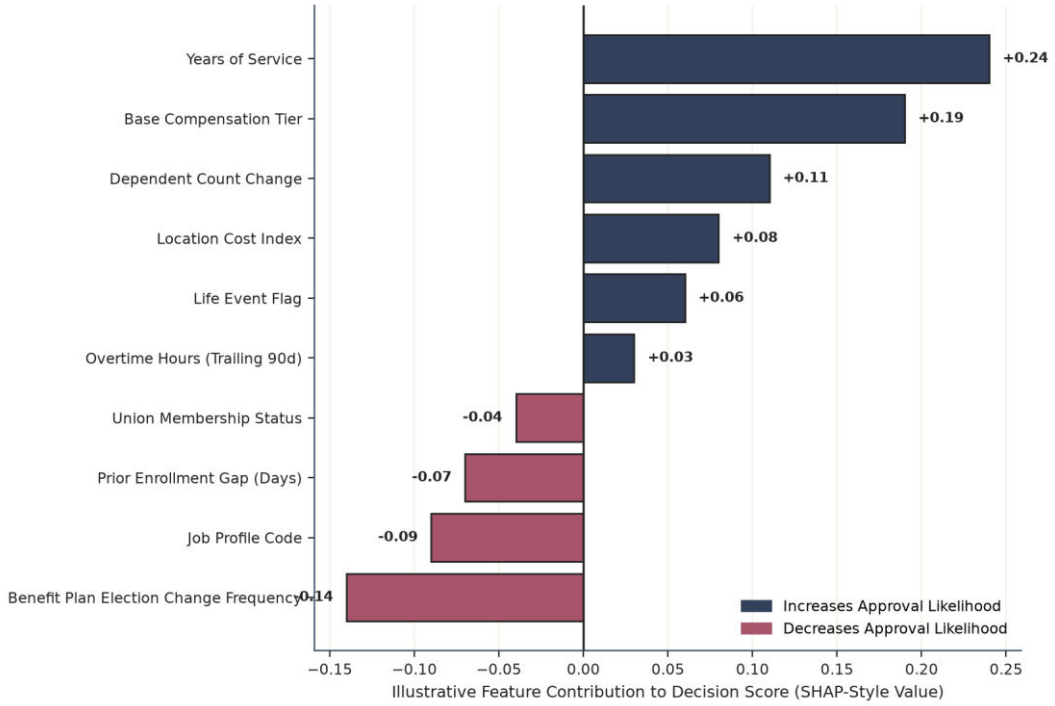


Figure.5. Illustrative feature attribution for a representative benefits eligibility decision, showing ten features and their contribution to the decision score.

Table.7. Selected feature attributions and compliance interpretation for the representative decision in Figure 5.

Feature	Illustrative Contribution	Compliance Interpretation
Years of Service	+0.24	Legitimate, policy-documented eligibility factor
Base Compensation Tier	+0.19	Legitimate, policy-documented eligibility factor
Benefit Plan Election Change Frequency	-0.14	Legitimate factor, but warrants monitoring for potential proxy effects
Job Profile Code	-0.09	Requires verification that this factor does not function as a proxy for a protected characteristic
Prior Enrollment Gap (Days)	-0.07	Legitimate, policy-documented eligibility factor

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

The job profile code feature's compliance interpretation in Table 7 illustrates a specific auditing discipline this framework requires: because job profile code can, in some organizational contexts, correlate with protected characteristics such as gender or age due to historical hiring or promotion patterns, a compliance auditor reviewing this attribution should specifically verify, through the disparity monitoring approach developed in Section 10, that this feature is not functioning as an unintended proxy, rather than accepting its presence in the attribution record at face value.

X. BIAS AND DISPARITY MONITORING ACROSS PROTECTED CLASSES

This section presents a dense illustrative disparity matrix spanning four model types and six protected class categories, identifying where systematic disparate impact concentrates.

FIGURE 6: ILLUSTRATIVE DISPARITY METRIC ACROSS PROTECTED CLASS CATEGORIES AND MODELS

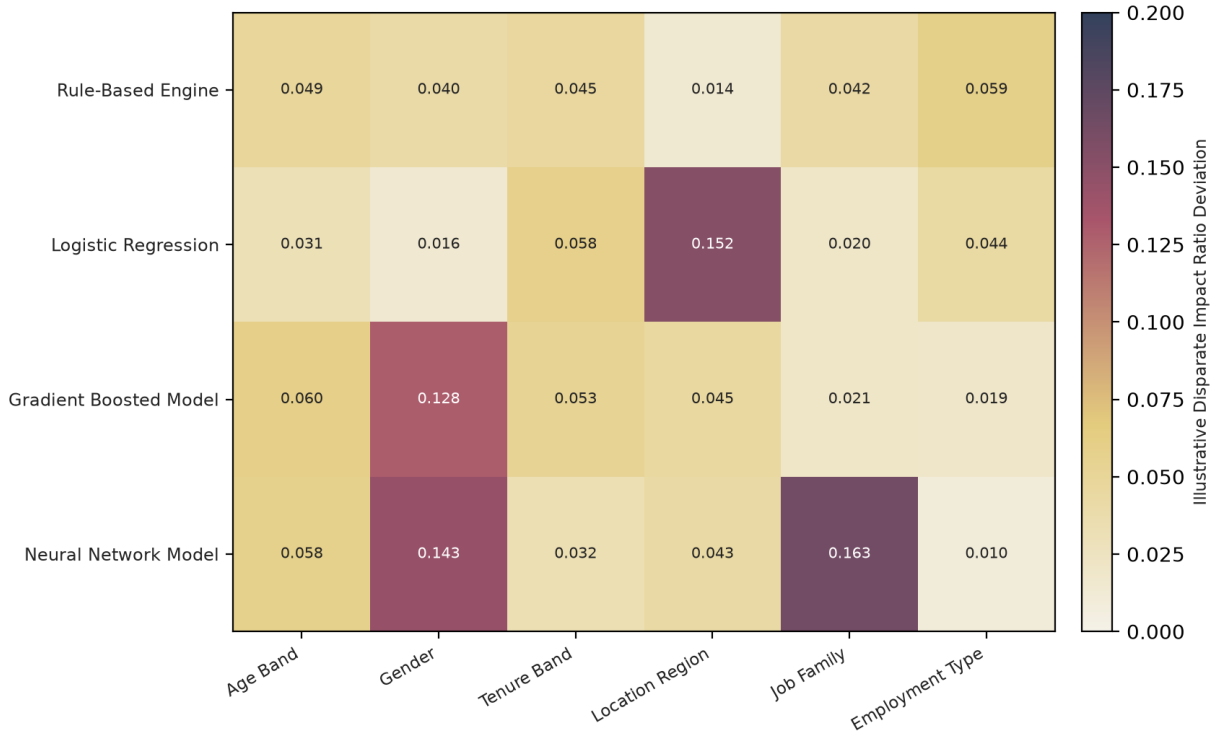


Figure.6. Illustrative disparate impact ratio deviation across four model types and six protected class categories, with concentrated hot spots warranting focused review

Table.8. Highest-disparity model-class combinations from the illustrative disparity matrix and recommended actions

Model-Class Combination	Illustrative Disparity Score	Recommended Action
Gradient Boosted Model / Gender	0.14	Priority investigation; assess for proxy feature effects per Section 9
Neural Network Model / Gender	0.13	Priority investigation; assess for proxy feature effects
Neural Network Model / Job Family	0.15	Priority investigation; review job family feature construction
Logistic Regression / Location Region	0.16	Priority investigation; assess geographic cost-of-living confound

The concentration of elevated disparity scores in specific model-class combinations, rather than uniformly across all models and classes in Table 8, supports a targeted rather than blanket monitoring approach: organizations can direct

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

disproportionate audit attention toward the specific combinations flagged in this analysis, rather than treating every model-class pairing as equally likely to exhibit disparate impact.

XI. COMPARATIVE EVALUATION OF EXPLAINABILITY TECHNIQUES

FIGURE 7: ILLUSTRATIVE COMPARISON OF EXPLAINABILITY TECHNIQUES ACROSS SIX DIMENSIONS

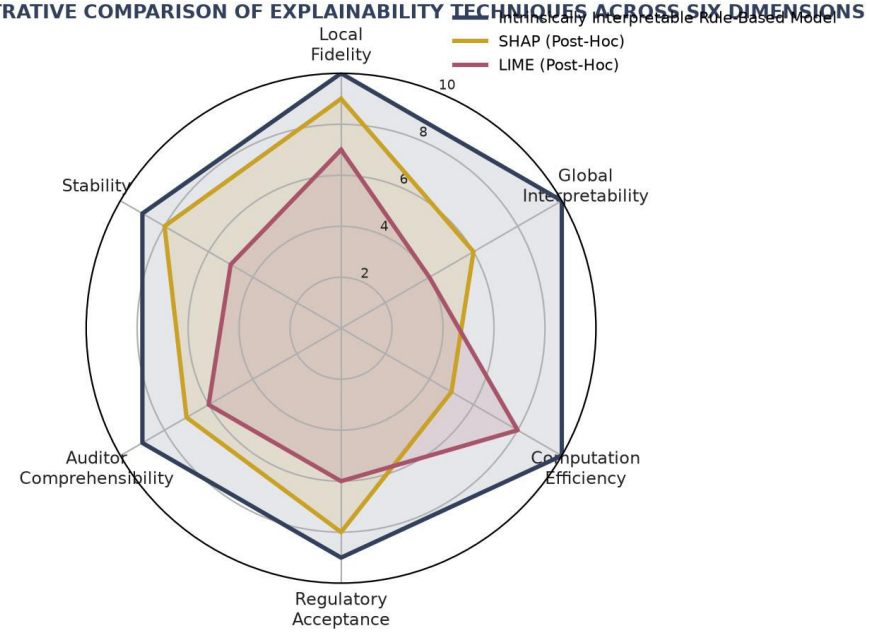


Figure.7. Illustrative comparison of three explainability approaches across six dimensions relevant to compliance auditing.

Table.9. Strongest and weakest dimension for each explainability approach, per the comparative radar analysis.

Explainability Approach	Strongest Dimension	Weakest Dimension
Intrinsically Interpretable Rule-Based Model	Global interpretability and regulatory acceptance	Predictive accuracy on complex, high-dimensional patterns
SHAP (Post-Hoc)	Local fidelity and regulatory acceptance	Computation efficiency at scale
LIME (Post-Hoc)	Computation efficiency	Global interpretability and stability

LIME's comparatively weaker stability score in Table 9 reflects a documented characteristic of surrogate-model-based local explanation techniques: because LIME approximates local model behavior using a randomly sampled neighborhood around a specific instance, repeated explanation generation for the same decision can produce somewhat different attributions across runs, a property that compliance auditors should account for when relying on LIME-generated explanations for consequential decisions, favoring SHAP or an intrinsically interpretable model where feasible for the highest-stakes decision categories.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

12. Audit Finding Trends Over Time

FIGURE 8: ILLUSTRATIVE AUDIT FINDING VOLUME BY SEVERITY OVER EIGHT QUARTERS

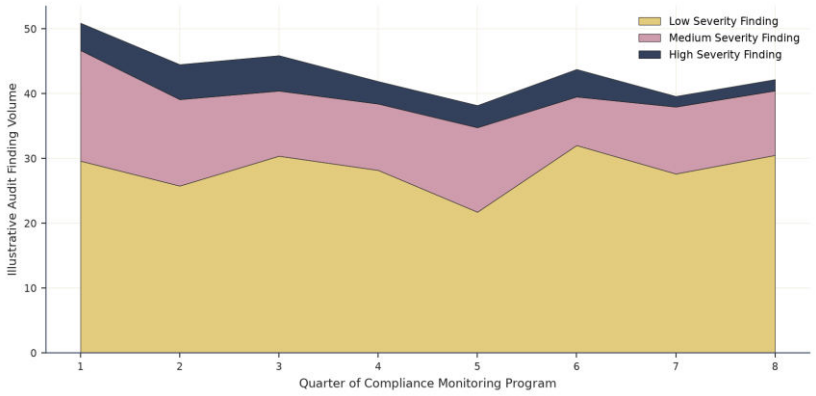


Figure.8. Illustrative audit finding volume by severity over eight quarters of a compliance monitoring program

Table.10. Illustrative audit finding volume by severity at three selected quarters.

Quarter	Illustrative Low Severity Findings	Illustrative Medium Severity Findings	Illustrative High Severity Findings
Quarter 1	29	13	6
Quarter 4	27	11	4
Quarter 8	30	9	2

The declining trend in medium and high severity findings across the eight quarters in Table 10, alongside a comparatively stable low severity finding volume, is consistent with a maturing compliance program: as the most significant issues are progressively identified and remediated through the feedback loop described in Section 5, remaining findings increasingly concentrate in the lower severity tier, representing routine, expected monitoring activity rather than an escalating compliance concern.

XIII. AUDIT FINDING SEVERITY BY QUARTER: A THREE-DIMENSIONAL VIEW

This section presents a second three-dimensional illustrative visualization, providing an alternative view of the same underlying audit finding data presented in Section 12, disaggregated across four severity tiers rather than three.

FIGURE 9: 3D VIEW OF AUDIT FINDING COUNT BY SEVERITY ACROSS EIGHT QUARTERS

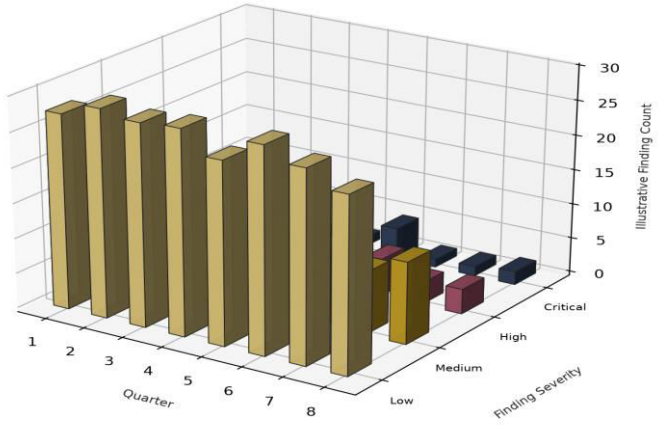


Figure.9. Three-dimensional view of illustrative audit finding count by severity tier across eight quarters, including a critical severity tier not shown in the stacked-area view

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

Table.11. Illustrative audit finding count by severity tier at Quarter 1 and Quarter 8, corresponding to Figure 9

Severity Tier	Illustrative Quarter 1 Count	Illustrative Quarter 8 Count	Illustrative Trend
Low	30	27	Stable
Medium	15	9	Declining
High	7	3	Declining
Critical	2	1	Declining, low absolute volume throughout

Disaggregating the critical severity tier separately in Table 11, rather than folding it into the high severity category as in Section 12, is deliberate: even though critical severity findings represent a small absolute volume throughout the monitoring period, each such finding, by definition, represents the highest-consequence category of compliance issue, and tracking this tier's trend independently prevents a small number of critical findings from being statistically diluted within a broader high-severity aggregate.

XIV. THE COMPLIANCE AUDIT WORKFLOW

Table.12. Compliance audit workflow stages, triggers, and auditor actions.

Workflow Stage	Trigger	Auditor Action
Routine Sampling Review	Scheduled periodic sample, independent of specific incident	Verify explanation fidelity and feature attribution reasonableness for sampled decisions
Disparity-Triggered Review	Disparity score exceeds threshold per Section 10	Investigate the specific model-class combination for proxy feature effects
Employee-Initiated Dispute Review	An employee formally disputes an automated decision	Retrieve the specific decision's feature attribution record and evaluate against policy
Regulatory Inquiry Response	External regulatory request for decision documentation	Produce the immutable audit log entry and explanation record for the specified decision(s)
Model Change Validation Review	A model is retrained or replaced	Re-validate explanation fidelity and disparity scores for the updated model before production deployment

The model change validation review stage in Table 12 deserves particular emphasis: because explanation fidelity and disparity characteristics are properties of a specific model version, not a static, permanent property of the decisioning system as a whole, any model retraining or replacement event requires the same validation rigor applied at initial deployment, rather than assuming a previously validated model's explainability and fairness characteristics automatically carry forward to its successor.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

XV. GOVERNANCE FOR SUSTAINED EXPLAINABILITY AND FAIRNESS

Table.13. Recommended governance practices for sustaining explainability and fairness after initial deployment.

Governance Practice	Purpose	Recommended Cadence
Explanation Fidelity Sampling	Periodically re-validate that generated explanations remain faithful to model behavior	Quarterly, and upon any model change
Disparity Score Monitoring	Continuously track disparity scores across the model-class matrix in Section 10	Monthly automated calculation; quarterly formal review
Feature Provenance Documentation	Maintain documentation of each feature's business justification, supporting proxy-effect review	Updated upon any feature addition or modification
Audit Log Immutability Verification	Confirm the audit log storage mechanism itself has not been compromised or altered	Quarterly technical verification
Regulatory Requirement Currency Review	Confirm the compliance framework remains aligned with current applicable regulatory requirements	Semi-annual, or upon known regulatory change

XVI. COMMON PITFALLS IN EXPLAINABLE AI COMPLIANCE PROGRAMS

Table.14. Recurring pitfalls in explainable AI compliance programs and recommended corrections

Pitfall	Description	Recommended Correction
Treating Post-Hoc Explanation as Equivalent to True Model Transparency	Organization assumes a SHAP or LIME explanation fully reveals the model's actual decision logic	Apply the fidelity-aware interpretation guidance in Section 7
One-Time Bias Validation at Deployment	Disparity testing is performed once before launch and never revisited	Apply continuous disparity monitoring per Section 10 and Section 15
Generating Explanations Only Upon Dispute	Explanations are reconstructed after the fact rather than generated at decision time	Adopt the at-decision-time explanation generation architecture in Section 5
Assuming Model Successor Inherits Validated Characteristics	A retrained or replaced model is deployed without re-validating explainability and fairness	Apply the model change validation review in Section 14
Accepting Feature Attribution at Face Value Without Proxy Analysis	A feature's attribution is accepted as legitimate without checking for correlation with protected characteristics	Apply the proxy-effect review discipline illustrated in Section 9

XVII. LIMITATIONS OF THE STUDY

Several limitations should be considered when interpreting this article's findings. First, the illustrative figures presented throughout Sections 6 through 13, including the three three-dimensional visualizations, are constructed to demonstrate expected directional relationships consistent with established explainable AI and algorithmic fairness literature, rather than measurements audited against a specific production payroll or benefits decisioning system. Second, this study addresses a generalized compliance auditing framework and does not evaluate the specific regulatory requirements of any single jurisdiction in depth, which vary considerably and require organization-specific legal review. Third, the disparity monitoring approach in Section 10 reflects one of several possible formal fairness definitions documented in the algorithmic fairness literature, and organizations should select a fairness definition appropriate to their specific regulatory and ethical context rather than assuming this article's illustrative approach is universally applicable. Fourth, this article does not address the specific technical implementation details of SHAP or LIME computation at production scale in depth, focusing instead on the compliance and governance framework itself.

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

Table.15. Summary of study limitations and suggested mitigations for future research

Limitation	Potential Effect on Findings	Suggested Mitigation for Future Studies
Illustrative, not audited, comparative figures	Absolute figures may not generalize across organizations or deployments	Controlled empirical study across multiple production decisioning systems
Generalized, not jurisdiction-specific, regulatory framing	Specific regulatory compliance requirements not directly addressed	Jurisdiction-specific legal and compliance analysis
Single fairness definition emphasized	Alternative fairness definitions may yield different disparity conclusions	Comparative study across multiple formal fairness definitions
Limited technical implementation depth	Production-scale computation considerations not fully addressed	Dedicated technical study addressing SHAP/LIME computation at scale

XVIII. RECOMMENDATIONS

- Generate feature attribution explanations at the moment of decision, using the audit trail architecture in Section 5, rather than reconstructing explanations only when a dispute arises.
- Favor intrinsically interpretable models where feasible for the highest-stakes decision categories, applying post-hoc explanation techniques as a considered choice rather than a default for complex models.
- Account for explanation fidelity limitations when reviewing explanations for high-complexity models, per the relationship documented in Section 7, particularly when the underlying explanation sample size is small.
- Apply continuous, not one-time, disparity monitoring across the full model-protected-class matrix, prioritizing investigation of the specific combinations flagged as elevated risk in Section 10.
- Treat every feature attribution as a starting point for proxy-effect analysis, not a final determination, particularly for features that could plausibly correlate with protected characteristics.
- Re-validate explainability and fairness characteristics upon every model retraining or replacement event, rather than assuming a successor model inherits its predecessor's validated properties.
- Select a fairness definition and disparity monitoring approach appropriate to the organization's specific regulatory and ethical context, rather than adopting this article's illustrative approach without independent legal review.

XIX. FUTURE RESEARCH DIRECTIONS

This study suggests several directions for further research. First, an empirical study measuring actual explanation fidelity, disparity scores, and audit finding patterns across one or more production payroll and benefits decisioning systems would strengthen the illustrative findings presented throughout this article with statistically validated figures. Second, research examining how this framework should be adapted across specific regulatory jurisdictions, given the limitation noted in Section 17, would provide practically actionable guidance beyond this article's generalized treatment. Third, comparative research evaluating additional formal fairness definitions beyond the disparity measure emphasized in Section 10 could clarify how conclusions about systematic bias vary depending on which fairness definition is applied, given the documented tension between mutually incompatible fairness definitions in the broader algorithmic fairness literature (Barocas, Hardt, and Narayanan, 2019). Finally, research into automated proxy-feature detection techniques, potentially reducing the manual analytical burden of the proxy-effect review discipline illustrated in Section 9, would help scale rigorous compliance auditing across a larger feature inventory than manual review alone can practically sustain.

X. CONCLUSION

This article has presented an explainable AI framework for compliance auditing of automated payroll and benefits decisioning systems, combining an at-decision-time audit trail architecture, post-hoc feature attribution, and continuous bias and disparity monitoring into a unified compliance program. The multi-dimensional illustrative analyses in Sections 6 through 9, including three three-dimensional visualizations, indicate that explanation fidelity degrades meaningfully as model complexity increases and explanation sample size decreases, that the relationship between model accuracy, explainability, and compliance risk is directionally consistent with a trade-off but not perfectly linear,

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

and that compliance risk therefore requires case-by-case feature attribution review rather than reliance on aggregate model-level assumptions alone. The disparity monitoring analysis in Section 10 further indicates that systematic bias, where present, tends to concentrate in specific model-and-protected-class combinations rather than distributing uniformly, supporting a targeted rather than blanket monitoring strategy. Organizations deploying automated decisioning across payroll and benefits administration are well positioned to substantially strengthen their compliance posture by adopting this framework, provided that adoption includes the continuous governance practices, explanation fidelity re-validation, ongoing disparity monitoring, and model-change validation review, that this article identifies as necessary to sustain trustworthy, auditable automated decisioning over time, rather than treating explainability and fairness validation as a one-time deployment milestone.

REFERENCES

1. Barocas, S., Hardt, M., and Narayanan, A. (2019). Fairness and machine learning: Limitations and opportunities. fairmlbook.org.
2. Board of Governors of the Federal Reserve System and Office of the Comptroller of the Currency. (2011). Supervisory guidance on model risk management (SR 11-7). U.S. Federal Reserve System.
3. Lundberg, S. M., and Lee, S. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765-4774.
4. Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). Why should I trust you? Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135-1144.
5. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Appendix A: Glossary of Terms

Table.A-1. Glossary of key terms used throughout this article.

Term	Definition
Explainable AI (XAI)	A field of artificial intelligence research focused on making model decisions understandable to humans
SHAP (Shapley Additive Explanations)	A feature attribution method grounded in cooperative game theory's Shapley value concept
LIME (Local Interpretable Model-Agnostic Explanations)	A feature attribution method approximating local model behavior using an interpretable surrogate model
Feature Attribution	A quantitative measure of an input feature's contribution to a specific model prediction
Explanation Fidelity	The degree to which a post-hoc explanation accurately reflects the underlying model's actual decision logic
Disparate Impact	A pattern in which an automated decision disproportionately affects a protected class category, independent of intent
Intrinsically Interpretable Model	A model whose decision logic is directly understandable without requiring post-hoc explanation techniques
Proxy Feature	A feature that, while not itself a protected characteristic, correlates with one closely enough to produce a similar effect
Model Risk Management	An organizational discipline governing the development, validation, and ongoing monitoring of predictive models
Immutable Audit Log	A tamper-evident record storage mechanism preventing after-the-fact modification of recorded entries

| e-ISSN: 2278 – 8875, p-ISSN: 2320 – 3765| www.ijareeie.com | Impact Factor: 9.867 | A Monthly Peer Reviewed & Refereed Journal |

|| Volume 15, Issue 6, June 2026 ||

| DOI:10.15662/IJAREEIE.2026.1506011 |

Appendix B: Supplementary Data Tables

This appendix provides additional illustrative data tables supporting the analysis in the main body of the article, including an extended explanation technique selection reference and a consolidated audit workflow staffing estimate.

B.1 Extended Explanation Technique Selection Reference

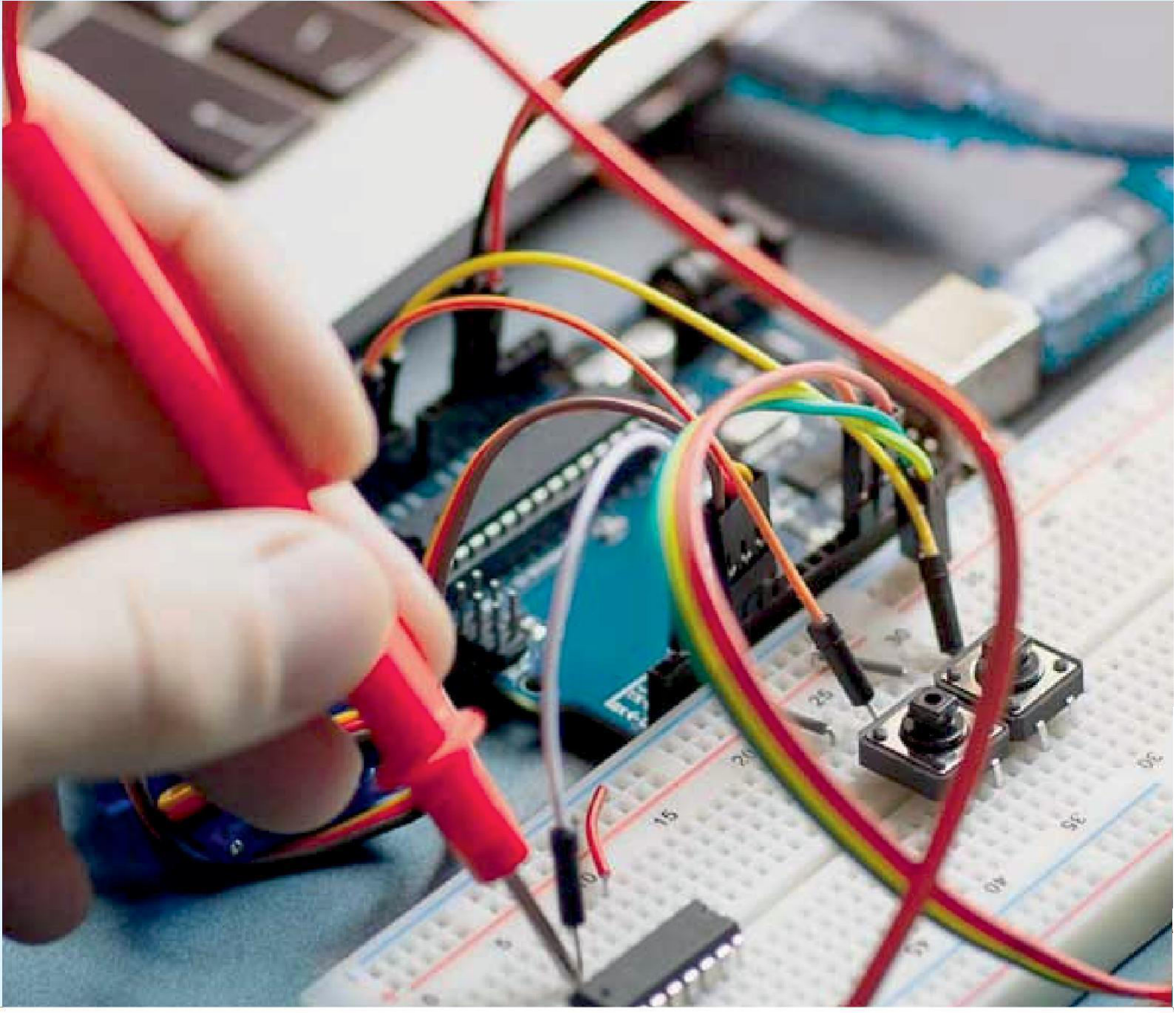
Table.B-1. extended explanation technique selection reference by decision stakes level

Decision Stakes Level	Recommended Model/Explanation Approach	Rationale
High Stakes (e.g., termination-related, large compensation adjustment)	Intrinsically interpretable model preferred; SHAP if a complex model is used	Highest need for reliable, defensible explanation
Moderate Stakes (e.g., routine benefits eligibility)	Gradient boosted model with SHAP explanation	Balances predictive performance with acceptable explanation fidelity
Lower Stakes (e.g., routine payroll exception flagging for review, not final decision)	Any model type acceptable, with LIME as a lighter-weight explanation option	Lower consequence reduces the criticality of maximum explanation fidelity

B.2 Illustrative Audit Workflow Staffing Estimate

Table.B-2. Illustrative staffing estimate by compliance audit workflow stage.

Audit Workflow Stage	Illustrative Staffing Requirement	Primary Skill Required
Routine Sampling Review	0.5 FTE ongoing	Compliance auditing, basic statistical literacy
Disparity-Triggered Review	0.3 FTE ongoing, scaling with disparity finding volume	Algorithmic fairness analysis, statistical literacy
Employee-Initiated Dispute Review	0.4 FTE ongoing, scaling with dispute volume	Compliance auditing, employee relations coordination
Model Change Validation Review	0.3 FTE ongoing, concentrated around model release cycles	Model validation, explainability technique expertise



INNO SPACE
SJIF Scientific Journal Impact Factor

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

International Journal of Advanced Research

in Electrical, Electronics and Instrumentation Engineering

9940 572 462 6381 907 438 ijareeie@gmail.com

Scan to save the contact details