THE COMPUTERTECH
(An International Peer Review Journal)
VOLUME 12; ISSUE 2 (JULY-DEC); (2026)
WEBSITE: THE COMPUTERTECH

Temporal Stability of Explainable AI in Financial Fraud Detection: A Framework for SHAP Auditing Under Drift

[1*] Ioana Baumann, [2] Lea Kovács
[1] Eötvös Loránd University, HUNGARY
[2] University of Belgrade, SERBIA

Abstract
Explainable artificial intelligence is increasingly used to support the review and governance of financial fraud models, yet a single explanation snapshot cannot by itself establish that a model remains interpretable, trustworthy, or reliable as transaction patterns evolve after deployment. This article develops a framework for auditing the temporal stability of feature-attribution explanations under concept drift and the severe class imbalance characteristic of fraud detection, and tests it empirically. It distinguishes local from global explanations, predictive attribution from causal explanation, and a genuine shift in explanation content from an unrelated shift in predictive performance, and it proposes a five-stage audit: a leakage-safe reference window, a documented and held-constant explainer configuration, comparison of explanation distributions across monitoring windows, triangulation against input drift and mature-label performance, and structured human review. The audit is implemented on the SHAP-guided, leakage-safe pipeline of Zuberi et al., applied to the public European credit-card dataset, with an XGBoost model and a logistic-regression baseline monitored across eight three-hour windows. Explanation change alone was an unreliable signal: global importance rankings for XGBoost fell well below their noise floor (Spearman 0.99) in windows 1, 2, 3, 4, but performance declined in only 3 of those, and a decline in window 8 (PR-AUC 0.20) occurred while the global ranking stayed stable (Spearman 0.90). Restricting the audit to confirmed fraud cases exposed that decline, with median attributions for V4 and V14 shifting by -5.2 and -2.2. The logistic baseline had more stable explanations and stronger day-two performance. Segment analysis by transaction amount showed recall gaps that aggregate figures concealed, although segment fraud counts were small. The results support triangulated and stratified explanation monitoring over explanation-only monitoring; the reviewer-based proposition could not be tested without human reviewers.

Keywords: Explainable AI; SHAP; Financial Fraud; Temporal Drift; Model Governance; Class Imbalance; Explanation Stability; Human Review

Introduction

Explainability has become a near-universal requirement in financial machine learning because a fraud score, however it is produced, can influence transaction blocking, customer-facing friction, analyst workload, and regulatory scrutiny, each of which creates a practical demand for some account of why the model reached the score it did. The presence of an explanation method attached to a model, however, does not by itself guarantee that the model is understandable, stable, or fair; a feature-attribution method explains a model's behavior relative to that specific model and the data

distribution it was fit on, and it does not, on its own terms, establish the underlying causal mechanism that produced an outcome. In fraud detection specifically, this distinction becomes materially more consequential once the transaction environment is allowed to change, because an explanation computed against an earlier population can continue to look plausible while silently describing a population that no longer exists.

Zuberi et al.[1] incorporate SHAP-guided feature selection within a leakage-safe, chronologically evaluated pipeline, a useful methodological foundation precisely because feature selection in their design is performed strictly on training data rather than on the complete dataset, avoiding the optimistic bias that would otherwise result from letting later information influence an earlier selection step. Their work, however, is concerned with feature selection at a single point in the modeling pipeline; it does not itself address whether the resulting explanations remain stable across time, nor how a change in those explanations relates to a change in the model's actual predictive performance. This article develops a framework for auditing explanations specifically under temporal drift and severe class imbalance, with its central concern being diagnostic validity, whether an observed explanation change is telling an analyst something true and actionable about the model, rather than mere visual interpretability, whether a chart of feature importances looks clean and plausible on its face.

Four objectives organize the remainder of this article: distinguishing predictive, attribution-based explanation from causal explanation, since the two are routinely conflated in applied practice; proposing concrete measures of explanation stability across successive temporal windows; connecting any observed explanation change to both input-level drift and mature, outcome-confirmed performance, so that an explanation shift is interpreted in context rather than in isolation; and specifying a governance protocol for structured human review of explanation evidence. The measures are then applied in an empirical audit of the Zuberi et al. pipeline on the public European credit-card dataset, so that the first two propositions can be tested on data; they are not yet validated instruments for deployment.

Explainability, Drift, and the Limits of a Single Snapshot

Local and Global Explanations

Lundberg and Lee [2] present SHapley Additive exPlanations as a unified feature-attribution approach grounded in cooperative game theory's Shapley-value framework, allocating credit for a given prediction across the input features that produced it in a manner that satisfies a specific, theoretically motivated set of consistency properties. Local explanations, in this framework, allocate contributions to one individual prediction and can support the review of a single flagged transaction; global summaries aggregate contributions across many predictions and can help an analyst identify which features the model relies on most heavily overall. Neither level of explanation, taken alone, establishes whether the underlying decision the model reached is actually justified; a locally plausible explanation can still rest on a model that is poorly calibrated or operating on a stale population, and a reassuring global summary can conceal substantial heterogeneity in how the model behaves on the specific, operationally important cases, such as borderline fraud, that a global average necessarily smooths over.

A later, tree-specific algorithmic refinement made exact SHAP computation practical at the latency and scale a production fraud-scoring system actually requires [3], which is part of why SHAP in

particular, rather than a more computationally expensive alternative, has become a default explainability layer for the gradient-boosted and random-forest models common in fraud detection. That computational tractability is a genuine practical advantage, but it does not change the more basic point that a SHAP value, however efficiently it is computed, remains an attribution relative to a specific model and background distribution rather than an independent measurement of the world.

Interpretability Claims and Their Limits

Lipton's widely cited caution about the mythos of model interpretability observes that the term is used to mean several different, not always compatible things, including transparency into a model's internal mechanics, post-hoc explanation of its outputs, and human simulability of its decision process, and that conflating these senses can make an explanation method appear more epistemically authoritative than it actually is [4]. A feature-importance ranking is not equivalent to a faithful causal account of why an outcome occurred; in a highly correlated feature space, attributed importance can be distributed across several correlated variables in ways that depend heavily on the specific explainer's own internal assumptions about how to divide credit among them, rather than on any fact about which feature is truly doing the causal work. Explanations should therefore be validated against the specific use they are being put to, the expertise of the person interpreting them, and the decision context in which they will be acted on, rather than treated as a generically reliable output that means the same thing in every context it is deployed in.

Temporal Non-Stationarity

Concept drift occurs when the relationship between input features and the outcome of interest changes over time, a phenomenon the broader non-stationary-learning literature treats as requiring explicit monitoring and adaptation rather than a one-time training exercise [5]. Under drift, an explanation generated against a model trained on an earlier period may no longer characterize either a model retrained on more recent data or the current transaction population the live model is actually scoring, which means an explanation's shelf life is itself an empirical question rather than something that can be assumed indefinitely. Explanation stability and predictive stability are related but importantly not identical: an explanation can change substantially while measured performance remains essentially stable, for instance because the underlying ranking function has not changed even though the correlation structure among its inputs has; conversely, an explanation can remain superficially similar while performance quietly deteriorates because of an unmonitored variable or a delayed-label artifact that the explanation method, by construction, has no way of surfacing. A closely related formalization, covariate shift adaptation, isolates the narrower case in which the input distribution moves while the underlying input-outcome relationship itself stays fixed [6], which is a useful limiting case for reasoning about when an explanation shift reflects a benign redistribution of inputs rather than a genuine change in what the model has learned to rely on.

Class Imbalance and Explanation Populations

When fraud is rare, which it is by a wide margin in virtually every production setting, a global explanation computed over the full transaction population will be dominated numerically by legitimate transactions, potentially obscuring exactly the attribution patterns that matter most for the rare positive class the model exists to catch. Conversely, an explanation computed only over the flagged population can overrepresent false positives relative to confirmed fraud, since in a

severely imbalanced setting the flagged population is frequently dominated by cases the model got wrong rather than cases it got right. A defensible audit should therefore report explanation summaries separately for confirmed fraud, legitimate transactions, false positives, false negatives, and operationally relevant customer or transaction segments, subject to privacy constraints and a documented minimum-sample threshold below which a stratum's summary is not reported at all. Resampling techniques developed specifically for severe class imbalance, such as the synthetic minority over-sampling approach [7], are frequently used upstream of model training in exactly this kind of setting, and any such resampling choice should itself be disclosed alongside an explanation audit, since a model trained on resampled data can produce attribution patterns that partly reflect the resampling procedure rather than the original class-conditional feature distributions alone. Precision-recall analysis helps contextualize the operational meaning of the flagged population that any stratified explanation audit is being computed over[8], since a flagged set that is mostly false positives calls for a different interpretation of its explanations than one that is mostly confirmed fraud.

Proposed Temporal Explanation Audit

The proposed audit contains five stages: establishing a leakage-safe reference model and reference explanation distribution; generating local and global explanations using a documented, held-constant explainer configuration; comparing explanation distributions across successive time windows; triangulating any observed explanation change against feature-level drift and mature, outcome-confirmed performance; and recording a structured, human-reviewed interpretation rather than an automated verdict. The protocol is designed specifically to avoid the single most common misreading of explanation monitoring in practice, treating a change in feature importance, by itself, as proof of a causal or adversarial change in the underlying fraud-generating process.

Reference and Monitoring Windows

All model fitting and feature selection must occur strictly within the designated training period, following the same chronological discipline that leakage-safe drift-evaluation methodology already establishes for predictive-performance testing [1]. A reference explanation distribution is then computed on a held-out chronological reference window that the model never saw during fitting, and subsequent monitoring windows are processed using the same feature definitions and the same explanation configuration wherever that is technically possible, so that any observed difference in explanations can be attributed to the data or the model rather than to an inadvertent change in how the explanations themselves were computed. If the underlying model itself changes between windows, whether through retraining, recalibration, or a feature-set revision, explanations from before and after that change should be compared with particular care, since any observed difference may reflect data change, model change, or some combination of the two that a naive before-after comparison cannot cleanly separate. An adaptive-windowing approach, which automatically adjusts the effective size of a monitoring window when a statistical test detects a significant change within it[9], offers one disciplined way to decide when a new reference window is actually warranted, rather than resetting the reference arbitrarily on a fixed calendar schedule.

Measuring Explanation Stability

Let $\varphi_j(x)$ denote the SHAP contribution of feature j for transaction x. For each monitoring window, an analyst may compare the distribution of φ_j against the reference distribution using

standardized mean differences, quantile-level changes, or a distributional-distance measure appropriate to the feature's scale. These statistics are descriptive rather than inferential in the strict sense, and they should always be reported together with their underlying sample sizes and an explicit measure of uncertainty; they should not be collapsed into a single, unvalidated composite explanation-drift score that discards the very stratification the Class Imbalance and Explanation Populations section argues is necessary to interpret them correctly.

$$\Delta_j = \text{median}(\varphi_j,\text{monitor}) - \text{median}(\varphi_j,\text{reference})$$

Triangulation

An observed explanation shift should always be examined alongside three further pieces of evidence: the feature's own raw distribution, to check whether the shift is simply tracking a change in the feature's inputs rather than in how the model weights it; the model's output score distribution, to check whether the shift coincides with any change in the model's overall behavior; and mature, outcome-confirmed performance metrics once labels become available. If SHAP contributions change while recall, precision, and calibration remain stable on mature labels, an institution may reasonably choose to retain the model in production while simply documenting the observed shift for the record. If, instead, an explanation change coincides with a genuine performance decline, a deeper review of data quality, feature semantics, and the model's underlying assumptions is warranted before any further action is taken. The practitioner literature on deployed credit-card fraud systems is directly relevant to this triangulation step, since it establishes that label-verification latency, the delay between a transaction occurring and its outcome being confirmed, is itself a structural feature of the evaluation problem rather than an incidental delay that can be safely ignored when deciding how much weight an explanation-based signal should carry before mature labels are actually available[10][11].

Human Review Protocol

Reviewers conducting this audit should receive the prediction itself, the relevant feature values that produced it, the explanation method used together with a plain statement of its limitations, a historical comparison against the reference window where one is available, and the eventual confirmed outcome once it is known. The review process should explicitly record whether the explanation was understandable to the reviewer, whether it supported a plausible investigation hypothesis the reviewer could act on, and whether that hypothesis was subsequently and independently confirmed, rather than simply recording whether the reviewer found the explanation agreeable at a glance. Reviewers should, as a matter of protocol rather than individual judgment, never treat an explanation as the sole basis for a consequential decision about a customer or a transaction, precisely because, as the Interpretability Claims and Their Limits section argues, an attribution is not itself a causal account.

Evaluation Design

An empirical study carried out under this framework should compare models across successive chronological windows and examine predictive performance and explanation stability jointly, rather than as two separately reported, disconnected analyses. Model candidates in such a study should include at least one comparatively interpretable baseline alongside one or more higher-capacity models, such as a random-forest or gradient-boosted ensemble[12], not because simpler

models are assumed to be preferable, but because the trade-off between predictive utility, explanation consistency, and operational usability is itself the object of study and cannot be quantified from a single model class alone.

Predictive Measures

Precision, recall, F-scores, the Matthews correlation coefficient, the area under the precision-recall curve, and calibration quality should all be reported under this evaluation design. The Matthews correlation coefficient summarizes all four confusion-matrix outcomes in a single, balanced figure that is markedly less distorted by severe class imbalance than accuracy or even F1[13], while precision-recall analysis is particularly informative under the rare-positive prevalence that characterizes fraud detection[8], and the receiver operating characteristic curve, interpreted alongside Fawcett's broader treatment of what it can and cannot establish under imbalance, remains useful specifically for comparing ranking quality independent of any one operating threshold[14]. These metrics should be stratified by time window and, wherever permitted by sample size and privacy constraints, by the operationally relevant segments introduced earlier, and confidence intervals around each reported figure are essential given how few confirmed fraud cases a given window will typically contain.

$$MCC = (TP \cdot TN - FP \cdot FN) / \sqrt{[(TP+FP)(TP+FN)(TN+FP)(TN+FN)]}$$

Explanation Measures

Candidate explanation-stability measures include the rank correlation of global feature-importance rankings across windows, distributional-distance measures between feature-attribution values, and local explanation consistency for comparable individual cases. Each of these measures has a specific limitation that an evaluator should disclose rather than paper over: rank correlation can remain high even while the underlying contribution magnitudes change substantially, since rank correlation is, by construction, insensitive to magnitude; distance metrics can be sensitive to the scale on which a given feature happens to be measured; and "comparable cases" for local-consistency comparison can be difficult to define cleanly once the underlying transaction population itself is changing, which is precisely the setting this framework is built to address. The tree-specific SHAP algorithm that made exact attribution computation practical at production scale[3] also makes repeated computation of these stability measures across many monitoring windows computationally tractable, which matters because a stability measure that is too expensive to compute routinely will not, in practice, be computed routinely. Researchers applying this framework should preregister their primary stability measure in advance and should specifically avoid the practice of computing several candidate measures and then reporting whichever one looks most favorable after the fact.

Audit Outcomes

Operational outcomes under this evaluation design should include the proportion of explanations that trained reviewers judge to have been useful, the level of agreement between independent reviewers examining the same case, the eventual investigation outcomes those explanations informed, and the frequency with which the audit process itself surfaces a previously undetected data-quality issue. These outcomes should be measured against a transparent, pre-specified protocol of the kind described in the Human Review Protocol section, rather than being inferred informally

from a visual inspection of a feature-importance plot, which is precisely the shortcut this framework is designed to replace with a documented and auditable process.

Analytical Framework and Propositions

Explanation stability and predictive-performance stability can move independently of one another, and the framework's practical value depends on being able to read the combination of the two correctly rather than defaulting to either extreme, treating every explanation change as alarming or treating every stable-looking global summary as reassuring. Table 1 sets out an illustrative mapping between the pattern of evidence a triangulated audit might observe and the interpretation that pattern most plausibly supports, together with the action that interpretation would warrant; the table is a conceptual aid for structuring the triangulation step. The empirical audit reported below observed three of its patterns: explanation change with stable performance (window 1), joint change (windows 2 to 4) and stable global explanations with declining performance (window 8).

Table 1. Illustrative mapping from joint explanation and performance evidence to interpretation and recommended action.

Explanation pattern	**Performance pattern (mature labels)**	**Likely interpretation**	**Recommended action**
Stable globally and by segment	Stable	Model and population are both behaving as expected	Continue routine monitoring; no special review needed
Shifted globally, stable by segment	Stable	Population mix has changed without a change in conditional behavior	Document the shift; retain the model; re-check at the next window
Stable globally, shifted in a high-risk segment	Stable in aggregate, not yet assessed by segment	Aggregate metrics are concealing a segment-specific change	Trigger segment-level performance review before concluding the model is unaffected
Shifted (any pattern)	Declined	Evidence consistent with a genuine change in the input-outcome relationship	Escalate to full diagnostic review of data quality, feature semantics, and retraining need

Testable Propositions

Three propositions follow from this framework. The first, P1, is that explanation changes combined with outcome-based performance metrics provide more useful diagnostic evidence, in the sense of correctly distinguishing benign from consequential shifts, than explanation changes considered in isolation. The empirical audit supports P1: explanation-only monitoring of the XGBoost model would have flagged 4 windows, of which 3 coincided with a genuine decline, and would have

missed the decline in window 8, whereas triangulation with mature labels classified every window correctly and stratification by confirmed fraud explained the window-8 decline. The second, P2, is that segment-level explanation audits identify genuine deterioration that aggregate, population-level metrics mask, a proposition motivated by the label-latency and realistic-modeling concerns raised in the practitioner fraud-detection literature[10][11]. The audit supports P2 descriptively: recall for small-amount transactions fell well below the aggregate figure in the weakest windows, although segment fraud counts were too small for formal testing. The third, P3, is that structured reviewer guidance of the kind specified in the Human Review Protocol section improves the reproducibility of explanation-based investigations, in the sense that independent reviewers reach more consistent conclusions under structured guidance than without it. P3 requires human reviewers working on matched cases and could not be tested in this study; it remains a proposition for future work.

Empirical Audit on the European Credit-Card Benchmark

Setup

The audit uses the public European credit-card dataset of 284,807 transactions over two days, 492 of them fraudulent[16]. Following Zuberi et al.[1], the first 80 percent of day one was the training period: standardisation, SHAP-guided feature selection (15 inputs, chosen by mean absolute SHAP value from a class-balanced gradient-boosting surrogate) and oversampling were all fitted there only. The last 20 percent of day one was the reference window, and day two was divided into eight three-hour monitoring windows. Two models were audited, as the Evaluation Design section recommends: XGBoost with Borderline-SMOTE[17][7] and an interpretable logistic-regression baseline with SMOTE. Thresholds were fixed at their F2-optimal values on the reference window. Explanations were computed with TreeSHAP for XGBoost and linear SHAP for the logistic model, with the explainer configuration held constant across windows. Each window's SHAP values were computed for all frauds and up to 6,000 sampled legitimate transactions, with weights restoring the population mix. Three stability measures were reported: the Spearman correlation between the reference and window rankings of global mean absolute SHAP; the Jaccard overlap of the top five features; and the largest scale-standardised Wasserstein distance between reference and window attribution distributions. The same measures were also computed for the confirmed-fraud and alerted populations separately. A noise floor was obtained by comparing the two halves of the reference window, where no drift is expected: Spearman 0.986 for XGBoost and 0.971 for the logistic model.

Explanation Stability and Performance

Table 2 reports performance and explanation stability by window. On the reference window, recall and PR-AUC were 0.82 and 0.76 for XGBoost and 0.85 and 0.80 for the logistic model. A window is treated as a performance decline when PR-AUC falls more than 0.15 below the reference value, and as an explanation change when the global rank correlation falls below 0.80, far below either model's noise floor.

Table 2. Predictive performance and explanation stability by day-two window (Spearman = rank correlation of global mean |SHAP| with the reference window).

Window	Frauds	XGB recall	XGB PR-AUC	XGB Spearman (all)	XGB Spearman (fraud)	XGB max W	LR recall	LR PR-AUC	LR Spearman (all)
1	48	0.77	0.81	0.79	0.78	5.0	0.90	0.89	0.96
2	21	0.67	0.50	0.54	0.57	8.1	0.95	0.94	0.92
3	10	0.10	0.18	0.63	0.52	12.1	0.80	0.81	0.94
4	17	0.59	0.55	0.67	0.86	7.8	0.76	0.68	0.97
5	26	0.77	0.63	0.89	0.87	1.1	0.77	0.73	0.96
6	37	0.84	0.74	0.87	0.94	0.7	0.92	0.82	0.96
7	40	0.78	0.73	0.85	0.86	0.8	0.80	0.80	0.97
8	12	0.50	0.20	0.90	0.89	0.6	0.67	0.35	0.98

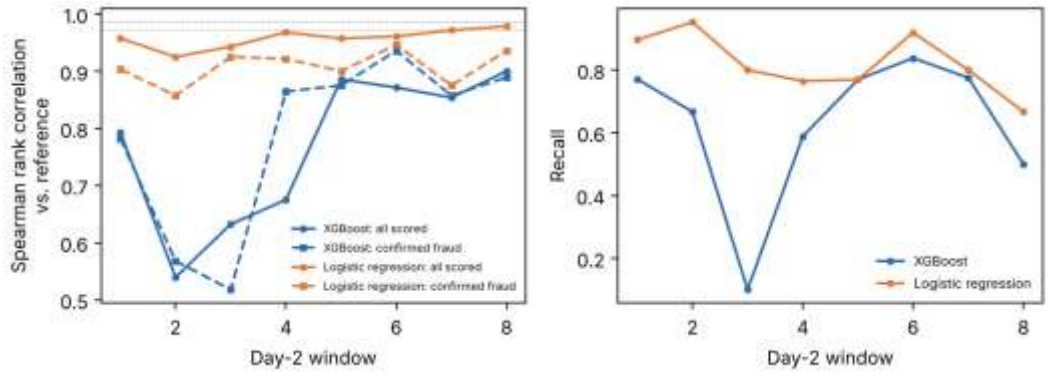


Figure 1. Rank stability of global SHAP importance (left; dotted lines are the reference-window noise floors) and recall (right) by day-two window [authors' analysis].

For XGBoost, explanation change and performance decline overlapped only partly. Global rankings fell below 0.80 in windows 1, 2, 3, 4, and PR-AUC declined in windows 2, 3, 4, 8. The two signals agreed in windows 2, 3, 4; window 1 showed explanation change with no performance decline, and window 8 showed a decline with no change in the global ranking. Explanation-only monitoring would therefore have produced one false alarm and one miss among four genuine declines. The early-morning windows, where explanation change was largest, had few transactions and an unusual mix of amounts, which shifted attributions without necessarily changing what the model had learned. The logistic baseline kept a stable ranking in every window (Spearman 0.92 to 0.98) and declined only in window 8. Its day-two PR-AUC exceeded that of XGBoost in 8 of 8 windows, so in this dataset the simpler model was both more stable to explain and more accurate after the reference period.

Triangulation and Stratified Explanation Populations

Window 8 shows why the Class Imbalance and Explanation Populations section argues for stratified reporting. Across all scored transactions the XGBoost ranking was stable (Spearman 0.90), and the median attribution change computed over the whole population was small for the leading features (V4: -0.36; V14: 0.20). Among confirmed frauds, however, the median attribution of V4 moved by -5.23 and that of V14 by -2.19, while their input PSI was only 0.24 and 0.12. The model was relying on its two leading features differently for the frauds it was now seeing, and this coincided with recall falling to 0.50. The population-level view hid this shift because legitimate transactions, more than 99 percent of every window, dominate any global summary. Triangulated against mature labels, the stratified signal pointed to a substantive change in fraud behaviour rather than a data artefact. Triangulated in the same way, window 1's explanation change, which came with stable performance, would have been documented and the model retained, as the Triangulation section recommends.

Segment-Level Audit

To test whether segment-level audits reveal deterioration that aggregate metrics mask, recall was computed separately for transactions below and at or above EUR 50, which accounted for 137 and 74 of the day-two frauds. In window 2, aggregate recall was 0.67, but 0.59 for small transactions. In window 8 it was 0.50 overall and 0.29 for small transactions (7 frauds). Small-amount fraud accounted for most of the deterioration that the aggregate figures averaged away. Because some windows contain only a handful of frauds per segment, these differences are descriptive and would need longer data to establish statistically.

Discussion and Implications

The audit proposed here reframes explainability as a monitored, ongoing component of model governance rather than a one-time feature-selection or reporting step completed once at launch and never revisited. SHAP and related attribution methods can meaningfully help analysts identify patterns worth investigating, but the resulting values remain dependent on the specific model, background distribution, and feature-correlation structure in force at the moment they were computed, and they do not, on their own, establish causality[3][4]. This distinction is particularly consequential whenever a model is retrained, since the same transaction can receive a materially different attribution purely because the model, its background distribution, or the correlation structure among its features has changed, independent of anything about the transaction itself having changed.

For researchers, the main implication is that explanation evaluation should be treated as inherently longitudinal rather than as a property established once and assumed to persist. A paper reporting explanation results should specify the explainer used, the background data it was computed against, the aggregation method applied, the class and segment composition of the population examined, and the comparison window used, with the same rigor that a predictive-performance claim would require. For financial institutions applying this framework operationally, explanation reports should be connected directly to investigation workflows rather than produced as a standalone artifact, and should communicate uncertainty explicitly; a reviewer needs to know both what the model attributed to a given feature and, just as importantly, what that attribution cannot prove, a point reinforced by the multi-institution field evidence showing how much practical difficulty surrounds

operationalizing any monitoring signal, explanation-based or otherwise, across real, independently governed institutions[15].

Conclusion

Temporal explanation auditing can meaningfully strengthen fraud-model governance when it is treated as one source of diagnostic evidence among several, rather than as either a one-time compliance artifact or an infallible early-warning signal in its own right. The protocol proposed in this article combines a leakage-safe reference window, a consistent and documented explanation-generation configuration, longitudinal comparison across monitoring windows, triangulation against both input drift and mature-label performance, and structured human review, and it is explicitly designed to avoid two opposite and equally damaging errors: assuming that any observed explanation change necessarily means the model has failed, and assuming that a stable-looking global explanation guarantees that the model's predictions remain reliable. Applied to the SHAP-guided pipeline of Zuberi et al. on a public benchmark, the audit showed that global explanation change alone both over-flags and under-flags, and that stratifying explanations by confirmed fraud recovered a performance decline that the global summary hid. Future work should test whether such audits improve investigation quality and reviewer consistency in real institutional settings, over longer periods than the 48 hours studied here. Until that evidence exists, the framework should be regarded as an empirically grounded but still partial proposal, offered to institutions and researchers who want to move explainability from a one-time feature-selection step toward a genuinely longitudinal component of model governance, rather than as a validated guarantee of explainability, fairness, or reliability in its own right.

References

[1] Zuberi, N., Rahman, M. M., Rashid, S. U., Ahsan, M. N., Rahman, M. D., & Mollah, M. H. O. R. (2026). A robust machine learning framework for detecting temporal drift in financial fraud prevention. *Scientific Reports*.

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

[3] Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, 2, 56-67.

[4] Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36-43.

[5] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), Article 44.

[6] Sugiyama, M., & Kawanabe, M. (2012). Machine learning in non-stationary environments: Introduction to covariate shift adaptation. MIT Press.

[7] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16, 321-357.

[8] Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLOS ONE, 10(3), Article e0118432.

[9] Bifet, A., & Gavaldà, R. (2007). Learning from time-changing data with adaptive windowing. In Proceedings of the 2007 SIAM International Conference on Data Mining (pp. 443-448).

[10] Dal Pozzolo, A., Caelen, O., Le Borgne, Y.-A., Waterschoot, S., & Bontempi, G. (2015). Learned lessons in credit card fraud detection from a practitioner perspective. Expert Systems with Applications, 41(10), 4915-4928.

[11] Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modeling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797.

[12] Breiman, L. (2001). Random forests. Machine Learning, 45, 5-32.

[13] Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genomics, 21, Article 6.

[14] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861-874.

[15] Rashid, S. U., & Sattiraju, S. R. (2025). Multi-Institution Validation of Drift-Resilient Early-Warning Models for SME Loan Distress and Transaction Fraud: Evidence from US Community Financial Institution Pilots. *International Journal of Modern Computing*, *8*(1), 40-54.

[16] Dal Pozzolo, A., Caelen, O., Johnson, R. A., & Bontempi, G. (2015). Calibrating probability with undersampling for unbalanced classification. In *2015 IEEE Symposium Series on Computational Intelligence* (pp. 159-166). IEEE.

[17] Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794).