Financial Technology and Innovation

♦ ISSN: 2836-5372 Vol: 06, No. 01, 2026 ♦ PP. 01–08
Article DOI: https://doi.org/10.54216/FinTech-I.060101



FULL LENGTH ARTICLE

Adaptive RegTech for Financial Complaint Operations: Temporal Institutional Risk Signals Outperform Text–Tabular Fusion in Predicting Untimely Responses

Samandarboy Sulaymanov[1,*] **Olimjonov Abbosjon Olimjonovich**[2]

[1] Tashkent State University of Economics, Tashkent, Uzbekistan
[2] Tashkent State University of Economics, Uzbekistan

Emails: sulaymanovsamandarboy@gmail.com · abbosolimjonovv@gmail.com

Received: December 11, 2025 Revised: February 04, 2026 Accepted: May 01, 2026 ⋆ Corresponding author

ABSTRACT

While the innovation in the financial sector can be measured by the products it offers customers, many valuable innovations are generated by operational technologies that facilitate the regulatory system to respond more quickly. This research designs and pilots an adaptive regulatory-technology approach to prioritize consumer complaints at high risk of an untimely institutional response. A leakage-safe chronological design is used to compare four approaches: static historical institutional-risk benchmark, short-horizon rolling-risk score, interpretable structured classifier, and text–tabular fusion classifier. Selection and calibration of models come before a non-biased one-month holdout period for evaluation. Untimely responses are very uncommon, and performance is evaluated based on precision–recall discrimination, calibration, and lift and recall under fixed review-capacity constraints, not just on accuracy. The seven-day rolling institutional-risk score proved to be the best performing operational group at a 2% review budget, identifying 75.7% of cases that arrive late with 9.9% accuracy, a 37.8-fold lift over random review. When the review budget was increased to 5%, 92.8% of cases that were untimely were captured. Unlike a common belief in financial NLP, incorporating any complaint-language indicator failed to improve financial ranking results: The fused model performed worse than any of the institutional-risk dynamic or static benchmarks. When service failures are concentrated in institutions, the findings indicate that the operational state near to the time of the failure could convey more information than richer complaint content. The suggested framework provides a clear low cost human-in-the-loop RegTech solution that allows to channel limited compliance focus and maintain auditability and chronological validity.

Keywords: Financial innovation ▪ RegTech ▪ Consumer complaints ▪ Explainable machine learning ▪ Rare-event prediction ▪ Operational risk ▪ Human-in-the-loop compliance

1. INTRODUCTION

Digital lending, alternative credit scoring, mobile payments, and platform-based intermediation are often considered the hallmarks of financial innovation. These developments have impacted the economics of screening, distribution and service delivery and have introduced new operational and conduct risk challenges [1, 2, 3]. Another, more hidden but increasingly significant field, is regulatory technology (RegTech): the use of data, analytics and automated controls to enhance the execution, monitoring and supervision responsiveness of compliance with regulations. In this area, the problem is not

© 2026 The Author(s). Published by ASPG. This article is licensed under CC BY 4.0.

so much a lack of data as the problem of what to do with the data that is available. The more challenging one is that of identifying where scarce human attention needs to be directed, before the failure of a service becomes a failure of regulation.

One of the most challenging contexts in which to provide decision support is consumer complaint handling. Financial institutions are provided with heterogeneous cases in all areas related to credit reporting, debt collection, banking, mortgage, lending, payments and virtual-currency services. The majority of cases are responded to within the time limits of the operation and a small minority of cases are responded to too late. The class-ratio is thus highly unfavourable. A model may have an apparently good hit-rate but be unable to find the critical cases. Furthermore, complaint operations are dynamic: fluctuations in staffing, process issues, vendor-related issues, system incidents or volume changes in cases can lead to clusters of untimely outcomes for limited periods of time. The institutional reputation score can be static and therefore fail to recognize persistent low performing institutions, or it can be too slow to respond to a new operational issue.

Modern finance is a field in which much recent research into financial machine learning has revealed the benefits of using nontraditional data for credit assessment and market access [4, 5, 6]. Meanwhile, the evidence of discrimination and imbalanced models effects has reinforced the call for governance, transparency and thoughtful validation [7, 8]. In financial applications, therefore, XAI has become a fundamental need [9, 10, 11]. Complaint triage is not the same thing, however, as credit scoring. The goal isn't to measure a person's potential creditworthiness; it's the risk that a financial institution's working process will not react in time. This distinction implies that the signal may be more important than the detail of a single complaint in determining recent institutional behaviour.

This paper develops an adaptive RegTech framework that ranks incoming complaints by the risk of an untimely response. The empirical design is deliberately chronological and audit-friendly. Historical institutional priors, rolling outcome rates, product and issue information, intake timing, and interpretable complaint-language indicators are constructed without access to future outcomes. Four models are evaluated on a sealed November–December 2024 test period after model specification and calibration using earlier months. The comparison is designed to answer three questions:

RQ1: Do dynamic institutional-risk signals improve rare-event ranking relative to static history and multivariable classifiers?

RQ2: How much of the untimely-response burden can be captured under realistic review-capacity constraints?

RQ3: Does adding complaint-language information provide incremental value beyond structured and temporal signals?

The study makes four contributions. First, it shifts financial-innovation analysis from customer acquisition and credit allocation toward regulatory service operations. Second, it introduces a prior-day, short-horizon institutional-risk signal that is simple enough to be audited and deployed without complex infrastructure. Third, it evaluates models through workload-sensitive operating points, which are more meaningful for compliance teams than accuracy or a single global threshold. Fourth, it reports a substantively important negative result: text–tabular fusion did not outperform simpler temporal risk signals. This result cautions against assuming that additional model complexity necessarily creates operational value.

2. LITERATURE REVIEW AND CONCEPTUAL BACKGROUND

2.1 Financial innovation and operational RegTech

FinTech changes the production and delivery of financial services by reducing search costs, automating decision processes, and exploiting digital traces [1, 2]. Digital footprints can improve screening where conventional credit information is limited [4], and technology-enabled lenders have expanded into consumer and small-business markets [5, 6]. These gains are accompanied by new dependencies on data pipelines, algorithms, third-party providers, and high-velocity operating models. The resulting risk landscape includes not only model risk but also service-level failure, delayed remediation, inconsistent communication, and inadequate complaint handling.

RegTech can be understood as an operational layer that converts regulatory obligations into observable controls, prioritized tasks, and evidence trails. Its value is therefore conditional on deployment design. A high-performing model that cannot be explained, monitored, or translated into action may have little compliance value. Conversely, a modest and transparent score can be useful if it consistently directs reviewers toward a concentrated pool of risky cases. This operational perspective is consistent with wider calls to align artificial intelligence with organizational processes, accountability structures, and public-policy objectives [12, 13].

2.2 Rare events, ranking, and limited review capacity

Untimely institutional responses are rare by design: the complaint system is expected to function, and most cases should be handled within the required window. This makes conventional accuracy unsuitable. For a prevalence below one percent, a classifier that always predicts a timely response will appear highly accurate while providing no operational benefit. Precision–recall analysis is better aligned with rare-event retrieval because it focuses on how many positive cases are found and how concentrated they are among reviewed cases. Evaluation should also reflect the actual decision process: a compliance unit may be able to review only 1%, 2%, or 5% of the daily queue.

The distinction between probability estimation and ranking is important. A model may rank cases effectively even if its raw probabilities require calibration. Conversely, a well-calibrated average risk estimate may not sufficiently concentrate rare events at the top of a queue. Classifier calibration should therefore be assessed separately from ranking performance [14]. Metric selection also needs to be robust to imbalance and aligned with the use case rather than chosen for convenience [15].

2.3 Explainability and text–tabular fusion in finance

Explainability has become central to responsible financial analytics because decisions affect customers, institutions, and regulators. The literature distinguishes intrinsically interpretable models from post-hoc explanations and emphasizes that explanation quality depends on the audience and task [16, 17, 18]. Local and global explanation methods can illuminate model behavior, but they do not remove the need for sound data design or chronological validation [19]. In financial risk management, explainable models support challenge, audit, and governance by connecting predictions to recognizable business variables [9, 10].

Text classification methods can extract useful signals from unstructured customer narratives [20]. Complaint language may reveal urgency, repeated contact attempts, disputed transactions, identity theft, legal escalation, or financial harm. Yet richer text does not guarantee better prediction. Narratives may be unavailable for many records, may describe customer harm rather than institutional processing capacity, and may contain vocabulary that is weakly connected to response timeliness. In addition, uncertainty can arise from both irreducible case variation and incomplete knowledge about changing operational conditions [21]. The present study therefore treats text fusion as an empirical proposition rather than an assumed improvement.

3. DATA AND METHODS

3.1 Data source and cohort construction

The analysis draws on records from the U.S. Consumer Financial Protection Bureau's Consumer Complaint Database, an official source that publishes complaint-level information after institutional processing and privacy controls [22]. The extract covers complaints received in California during 2024. Records were retained when the response-timeliness field contained a valid "Yes" or "No" value, and duplicate complaint identifiers were removed. The resulting cohort contained 268,699 distinct complaints. There were 787 untimely responses, corresponding to a prevalence of 0.293%.

The database includes the date received, product, issue, company, submission channel, complaint narrative when publication consent and disclosure conditions are met, and the institutional response fields. Narrative text was available for 80,737 records. The analysis uses only fields that can be associated with the complaint at intake plus historical outcomes that would have been known before the current complaint date. No company response text, closure description, or later consumer-dispute outcome was used as a predictor.

Table 1 reports the chronological partitions. January through August served as the initial development period, September and October as the validation and calibration period, and November and December as the independent test period. The final test set contained 58,120 complaints, including 152 untimely responses.

Figure 1 shows monthly volume and the untimely-response rate. The coexistence of changing complaint volume and a low, nonconstant event rate motivates chronological evaluation rather than random cross-validation.

Table 1. Cohort profile and chronological partitions

Cohort	Complaints	Untimely	Rate (%)
Full 2024 cohort	268,699	787	0.293
Development: January–August	156,819	506	0.323
Validation: September–October	53,760	129	0.240
Independent test: November–December	58,120	152	0.262
Narratives available	80,737	414	0.513

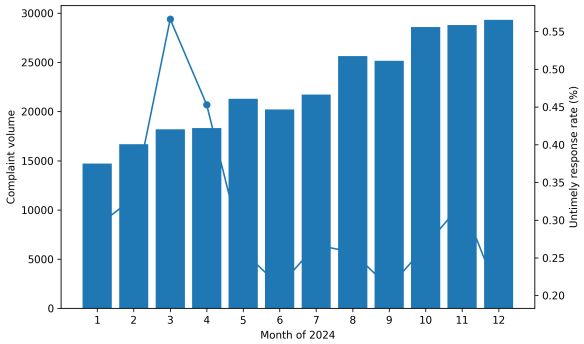


Figure 1. Monthly complaint volume and untimely-response rate during 2024.

3.2 Outcome and predictor construction

The binary outcome equals one when the field *Timely response?* is "No" and zero when it is "Yes." Predictor construction followed three principles: operational availability, chronological integrity, and auditability.

First, calendar variables included weekday and day of year. Second, product, issue, and company were represented through smoothed historical risk and relative frequency. For grouping variable g, the empirical-Bayes risk estimate was

$$\widehat{p}_g = \frac{y_g + \lambda \bar{p}}{n_g + \lambda}, \tag{1}$$

where y_g is the number of prior untimely outcomes in group g, n_g is its prior complaint count, $\bar{p}$ is the development-period prevalence, and $\lambda = 20$ controls shrinkage toward the overall rate. Shrinkage limits unstable estimates for small institutions or uncommon issues.

Third, dynamic institutional features counted prior untimely outcomes and prior complaints over 7-, 30-, 60-, and 90-day windows. For complaint i received on day t_i, the rolling rate for company c and window w was

$$R_{i,c}^{(w)} = \frac{Y_{c,[t_i-w,t_i)} + \alpha \bar{p}}{N_{c,[t_i-w,t_i)} + \alpha}, \tag{2}$$

where the half-open interval excludes the current day and every future observation. The smoothing constant was $\alpha = 5$. Excluding the current day prevents labels from other complaints processed on the same date from leaking into the prediction.

The fused specification added narrative length, word count, narrative availability, and indicators for interpretable linguistic themes: fraud, identity theft, disputes, payments or refunds, delay, prior contact, response attempts, and legal escalation. These indicators were chosen to preserve auditability and avoid an opaque, high-dimensional language model in a setting with substantial narrative missingness.

3.3 Models

Four approaches were compared.

1. **Historical risk**: the empirical-Bayes company risk estimated from all data available before the test period.
2. **Rolling risk**: the smoothed seven-day company untimely-response rate calculated from prior days only.
3. **Structured classifier**: an elastic-net stochastic-gradient logistic classifier using timing, rolling counts and rates, and smoothed company, product, and issue features.
4. **Fused classifier**: the structured classifier augmented with narrative length and interpretable language indicators.

The multivariable classifiers used standardized numerical inputs, class-balanced training weights, a logistic loss, elastic-net regularization, and averaged stochastic-gradient updates. Hyperparameters were fixed after evaluation on the September–October validation period. The model specifications were then refitted on January–October data. Raw probabilities from the structured and fused classifiers were calibrated through Platt scaling estimated only on validation predictions from the earlier January–August models. This separation ensures that the November–December test outcomes were not used for feature engineering, specification choice, calibration, or threshold selection.

3.4 Evaluation framework

Discrimination was evaluated with the area under the receiver operating characteristic curve (ROC AUC) and average precision, reported as PR AUC. Calibration was measured with the Brier score. Because the practical objective is to prioritize a fixed number of cases, operating points were computed at review budgets of 0.5%, 1%, 2%, 3%, 5%, and 10% of the queue.

For budget b, the top $\lceil bn \rceil$ complaints by score were selected. Recall, precision, and lift were then calculated as

$$\text{Recall}_b = \frac{TP_b}{P}, \tag{3}$$

$$\text{Precision}_b = \frac{TP_b}{\lceil bn \rceil}, \tag{4}$$

$$\text{Lift}_b = \frac{\text{Precision}_b}{P/n}, \tag{5}$$

where TP_b is the number of untimely cases found within the review budget, P is the total number of untimely cases, and n is the test-set size. Uncertainty for the best 2% operating point was estimated with 250 nonparametric bootstrap samples. All analysis was implemented in Python using pandas, NumPy, scikit-learn, and Matplotlib. The supplied code reconstructs the chronological features, models, tables, and PNG figures.

4. RESULTS

4.1 Complaint composition and operational heterogeneity

The data were dominated by credit-reporting complaints, which represented 227,073 records but had a relatively low untimely-response rate of 0.072%. The highest rates among major product groups occurred in prepaid cards (5.252%), student loans (5.204%), payday/title/personal/advance loans (4.695%), and money-transfer or virtual-currency services (2.863%). Debt collection generated the largest absolute number of untimely cases (236) among the listed products. Table 2 illustrates why both volume and conditional risk must be considered: a low-rate high-volume product may contribute many cases, while a smaller product can have a much higher local probability of delay.

Table 2. Complaint and untimely-response profile by leading product categories

Product	Complaints	Untimely	Rate (%)
Credit reporting or other personal consumer reports	227,073	164	0.072
Debt collection	13,924	236	1.695
Credit card	9,301	15	0.161
Checking or savings account	7,973	37	0.464
Mortgage	2,691	34	1.263
Money transfer, virtual currency, or money service	2,340	67	2.863
Student loan	1,518	79	5.204
Prepaid card	1,466	77	5.252
Vehicle loan or lease	1,319	17	1.289
Payday, title, personal, or advance loan	852	40	4.695

Institutional risk was even more concentrated. Several companies experienced short periods with a high proportion of untimely responses, while large institutions typically showed low unconditional rates. This pattern is consistent with operational episodes rather than a uniformly distributed failure process and supports the use of recent institutional history.

4.2 Out-of-time discrimination and calibration

Table 3 reports the untouched test results. The static historical-risk score achieved the highest PR AUC (0.291) and ROC AUC (0.987). The seven-day rolling-risk score followed closely in overall discrimination, with a PR AUC of 0.241 and ROC AUC of 0.986, while attaining the lowest Brier score (0.002299). The structured classifier achieved a PR AUC of 0.078, and the fused classifier achieved 0.068. Both remained far above random ranking in the high-recall region, but neither matched the simple institutional benchmarks.

Table 3. Performance on the independent November–December test period

Model	ROC AUC	PR AUC	Brier score
Historical risk	0.987	0.291	0.002351
Rolling risk	0.986	0.241	**0.002299**
Structured classifier	0.981	0.078	0.002452
Text–tabular fused classifier	0.975	0.068	0.002452

The precision–recall curves in Figure 2 show that the historical score was particularly strong at the smallest review

fractions, while the rolling score became more competitive as the queue expanded. The distinction matters because global PR AUC averages performance over all thresholds, whereas compliance teams operate at a selected capacity.

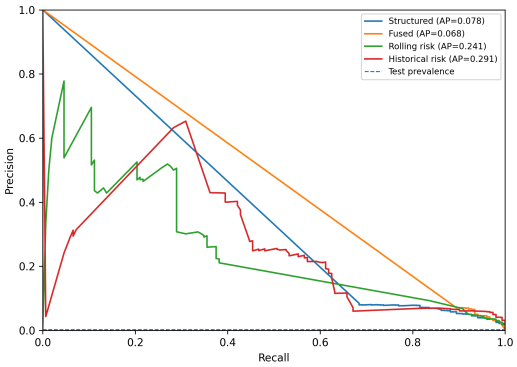


Figure 2. Precision–recall curves on the independent test period. AP denotes average precision.

Figure 3 compares observed and predicted rates across score groups. The event prevalence is low enough that small absolute deviations can appear large on a percentage scale. The rolling score provided the best Brier performance, indicating that its smoothed recent-history estimate was useful not only for ranking but also as a compact probability proxy. The multivariable models were Platt-calibrated; nevertheless, their ranking weakness limited operational utility.

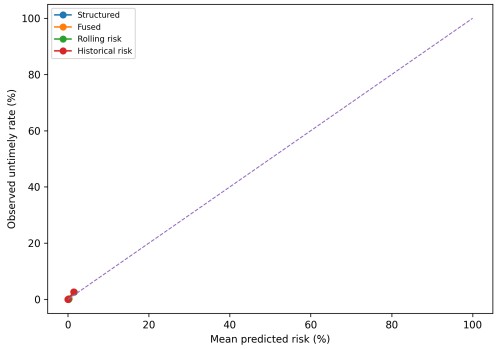


Figure 3. Calibration by quantile of predicted risk on the independent test period.

4.3 Review-capacity performance

Table 4 presents the central operational result. At a 0.5% review budget, the historical-risk benchmark identified 73 of 152 untimely cases (48.0%) with precision of 25.1%. At a 2% budget, the rolling-risk score identified 115 cases (75.7%), compared with 101 for historical risk, 98 for the structured classifier, and 79 for the fused classifier. Reviewing the top 1,163 rolling-risk cases therefore increased the untimely prevalence from 0.262% in the full test queue to 9.888%, a lift of 37.8.

The workload curves in Figure 4 clarify that there is no single model winner at every capacity. Historical risk is most effective for extremely narrow interventions, reflecting persistent high-risk institutions. Rolling risk is strongest around the practically important 2–3% region, where it detects emerging institutional episodes not fully summarized by long-run history. At larger budgets, the models converge because most positive cases have already entered the review set.

Table 4. Operational performance under fixed review budgets

Budget	Model	Reviewed	True positives	Recall (%)	Precision (%)
	Historical risk	291	73	48.0	25.1
	Rolling risk	291	59	38.8	20.3
	Structured	291	28	18.4	9.6
	Fused	291	20	13.2	6.9
	Historical risk	1,163	101	66.4	8.7
	Rolling risk	**1,163**	**115**	**75.7**	**9.9**
	Structured	1,163	98	64.5	8.4
	Fused	1,163	79	52.0	6.8
	Historical risk	2,906	148	97.4	5.1
	Rolling risk	2,906	141	92.8	4.9
	Structured	2,906	141	92.8	4.9
	Fused	2,906	143	94.1	4.9

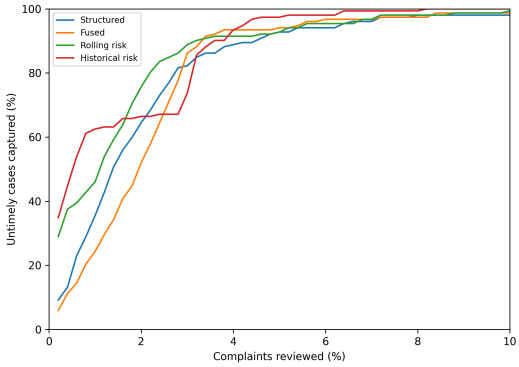


Figure 4. Share of untimely responses captured as the review budget increases.

Bootstrap analysis of the rolling-risk score produced a ROC AUC estimate of 0.986 (95% CI: 0.982–0.989) and PR AUC of 0.241 (95% CI: 0.175–0.316). At the 2% budget, recall was 0.757 (95% CI: 0.671–0.817) and precision was 0.099 (95% CI: 0.079–0.118). The intervals confirm strong concentration despite uncertainty arising from only 152 positive test cases.

4.4 Why did text fusion underperform?

The fused classifier's standardized coefficients are shown in Figure 5. Institutional exposure and history dominated the specification. Company frequency had the largest absolute coefficient, followed by recent complaint counts, the smoothed company prior, and positive counts over 30–90 days. Only a small number of language indicators appeared among the largest coefficients, and their contribution did not translate into better out-of-time ranking.

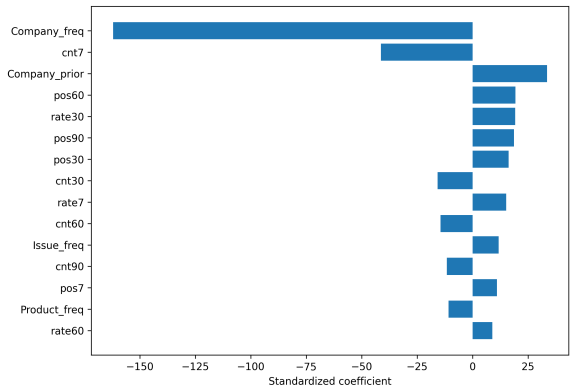


Figure 5. Largest standardized coefficients in the fused logistic classifier. Positive values increase the estimated log-odds of an untimely response.

Three mechanisms can explain this result. First, response timeliness is an institutional processing outcome; complaint language primarily describes the consumer's experience and may be only indirectly related to queue management. Second, narratives were available for 30.0% of the cohort, creating an uneven information channel. Third, temporal operational failures can be abrupt. A recent cluster of delayed responses may be more diagnostic than semantic similarity to historical complaints. The negative result does not imply that text is unhelpful for every complaint-management task. Text may be valuable for routing, topic discovery, severity assessment, or remediation design. It simply did not add incremental value for the narrow outcome studied here.

5. DISCUSSION

5.1 Temporal institutional state as the key operational signal

The main finding is that a transparent seven-day institutional-risk score can outperform a more elaborate text–tabular model at a realistic review capacity. This finding complements, rather than contradicts, the broader FinTech literature. Alternative data are valuable when they reveal latent borrower characteristics or reduce information asymmetry [4, 5]. In complaint operations, however, the target is generated by the institution's current process. Once the recent operational state is observed, additional information about the individual complaint may contribute little.

The contrast between static and rolling history is also instructive. Static historical risk achieved the highest aggregate PR AUC and dominated at the 0.5% and 1% budgets. It identifies institutions with persistent or repeated problems. The rolling signal performed better at the 2% and 3% budgets, where a compliance team can address both chronic and newly emerging cases. The two signals therefore represent different operational concepts: structural risk and transient state. A practical system could display both rather than force a single composite score.

5.2 A deployable human-in-the-loop architecture

The results support a lightweight deployment architecture. Each day, the system would update prior-day complaint and untimely-response counts by institution. New complaints would receive a rolling score and a static historical score. A configurable review budget would determine the number of cases placed in an enhanced-monitoring queue. Reviewers could then examine the complaint context, verify institutional capacity, request an internal status update, or initiate escalation. The score should not automatically determine regulatory liability or customer remedy.

This design has several governance advantages. The score is reproducible from a small set of counts; the time boundary is explicit; every prioritized case can be traced to recent institutional outcomes; and capacity can be adjusted without retraining a complex model. These properties reflect the principle that interpretable systems can be preferable in high-stakes settings when their performance is competitive [18]. They also facilitate model-risk management, because monitoring can focus on prevalence, drift in institutional volumes, window stability, and workload yield rather than on a large and opaque feature space.

5.3 Implications for financial innovation

The findings broaden the meaning of innovation in the financial sector. Innovation is not limited to new products or automated customer decisions. It can also consist of better operational sensing: detecting when an existing process is beginning to fail and allocating human capacity before the failure propagates. Such systems may reduce delayed responses, improve regulatory evidence, and protect consumers without automating the substantive adjudication of complaints.

For financial institutions, the workload results provide a concrete planning metric. A 2% enhanced-review queue concentrated approximately three quarters of untimely cases in the test period. This does not mean that every selected complaint would become untimely; precision was 9.9%, so human review remains necessary. The operational benefit is concentration. Instead of treating every case identically, the institution can direct limited specialist attention to a queue with a 37.8-fold higher event rate than the full population.

For regulators, the framework offers a way to monitor service responsiveness without relying solely on annual institutional averages. Short-horizon signals can reveal an emerging incident earlier, while static priors provide a view of persistent risk. For technology vendors, the findings caution against selling complexity as innovation. A transparent signal derived from clean event history may create more value than a richer natural-language model, particularly when the outcome is determined by back-office capacity rather than customer wording.

6. ROBUSTNESS, LIMITATIONS, AND FUTURE RESEARCH

Several design choices strengthen the analysis. The test period is strictly later than training and validation; rolling features exclude the current day; probability calibration is estimated before the final test; and operating points are reported across multiple capacity levels. Bootstrap intervals quantify sampling uncertainty for the principal result. The complete data extract, code, tables, and figures are supplied to permit independent reproduction.

The study also has limitations. First, the data describe complaints received in one U.S. state during one year. California provides a large and heterogeneous financial market, but institutional behavior and regulatory workflows may differ across jurisdictions. Multi-state and multi-year replication is needed. Second, the database is not a census of all consumer problems. Complaint submission depends on awareness, access, and individual choice, and published records are subject to the database's scope and disclosure rules [22, 23]. The results therefore concern prioritization within recorded complaints, not population-level consumer harm.

Third, the outcome indicates whether the institutional response was timely; it does not measure response quality, consumer satisfaction, or the adequacy of remediation. A timely response may still be unsatisfactory. Future work should develop multi-objective systems that jointly consider timeliness, dispute continuation, monetary relief, response quality, and fairness. Fourth, the model uses observed historical outcomes

as signals. A sudden policy or reporting change could alter their meaning. Deployment should include drift monitoring, periodic recalibration, and a documented override process.

Fifth, text features were intentionally interpretable and compact. More advanced language models might extract different signals, but they would introduce additional governance, privacy, computational, and reproducibility concerns. A useful next step is to compare domain-adapted language embeddings with the same sealed temporal design and to test whether they improve routing or severity assessment even when they do not improve timeliness prediction. Finally, causal interpretation is not warranted. A high rolling score indicates elevated recent risk; it does not identify the cause of the operational episode. Root-cause investigation remains a human and organizational task.

7. CONCLUSION

This study illustrates how simple, adaptive and auditable risk signals from financial innovations can enhance regulatory service operations. In a setting with very imbalanced complaints, a seven-day prior-only institutional-risk score identified 75.7% of the untimely cases by looking at 2% of the test queue, and reached a 37.8-fold lift compared to random selection. At 5% budget it captured 92.8% of untimely cases. Static institutional history still had value at very small review budgets, reaffirming the importance of separating chronic and transient risk.

The text–tabular classifier did not improve the rank. This negative finding is of practical importance: more data and more complex models don't necessarily mean better compliance technology. For outcomes that arise from institutional operations, the most relevant information may be the recent state of the institution. Therefore, a human-in-the-loop RegTech system based on transparent rolling signals can provide a credible way to earlier intervention, better allocation of compliance resources, and more accountable financial-service innovation.

DATA AVAILABILITY

The study data are included in the accompanying reproducibility package. The source extract is derived from the Consumer Financial Protection Bureau's Consumer Complaint Database. The package also provides the exact filtering and analysis code.

CODE AVAILABILITY

All code required to reconstruct the analytical cohort, chronological features, models, result tables, and PNG figures is included in the accompanying package.

ETHICS STATEMENT

The analysis uses de-identified public administrative records. No attempt was made to identify complainants, and no individual-level automated decision is proposed. The intended use is operational prioritization with human review.

REFERENCES

[1] A. V. Thakor, "Fintech and banking: What do we know?" *Journal of Financial Intermediation*, vol. 41, p. 100833, 2020.

[2] A. W. A. Boot, P. Hoffmann, L. Laeven, and L. Ratnovski, "Fintech: What's old, what's new?" *Journal of Financial Stability*, vol. 53, p. 100836, 2021.

[3] V. Murinde, E. Rizopoulos, and M. Zachariadis, "The impact of the FinTech revolution on the future of banking: Opportunities and risks," *International Review of Financial Analysis*, vol. 81, p. 102103, 2022.

[4] T. Berg, V. Burg, A. Gombović, and M. Puri, "On the rise of FinTechs: Credit scoring using digital footprints," *The Review of Financial Studies*, vol. 33, no. 7, pp. 2845–2897, 2020.

[5] M. Di Maggio and V. Yao, "Fintech borrowers: Lax screening or cream-skimming?" *The Review of Financial Studies*, vol. 34, no. 10, pp. 4565–4618, 2021.

[6] M. Gopal and P. Schnabl, "The rise of finance companies and FinTech lenders in small business lending," *The Review of Financial Studies*, vol. 35, no. 11, pp. 4859–4901, 2022.

[7] R. Bartlett, A. Morse, R. Stanton, and N. Wallace, "Consumer-lending discrimination in the FinTech era," *Journal of Financial Economics*, vol. 143, no. 1, pp. 30–56, 2022.

[8] A. Fuster, P. Goldsmith-Pinkham, T. Ramadorai, and A. Walther, "Predictably unequal? the effects of machine learning on credit markets," *The Journal of Finance*, vol. 77, no. 1, pp. 5–47, 2022.

[9] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable AI in fintech risk management," *Frontiers in Artificial Intelligence*, vol. 3, p. 26, 2020.

[10] N. Bussmann, P. Giudici, D. Marinelli, and J. Papenbrock, "Explainable machine learning in credit risk management," *Computational Economics*, vol. 57, no. 1, pp. 203–216, 2021.

[11] J. Černevičienė and A. Kabašinskas, "Explainable artificial intelligence (XAI) in finance: A systematic literature review," *Artificial Intelligence Review*, vol. 57, p. 216, 2024.

[12] Y. K. Dwivedi, L. Hughes, E. Ismagilova, G. Aarts, C. Coombs, T. Crick, Y. Duan, R. Dwivedi, J. Edwards, A. Eirug, V. Galanos, P. V. Ilavarasan, M. Janssen, P. Jones, A. K. Kar, H. Kizgin, B. Kronemann, B. Lal, B. Lucini, R. Medaglia, K. Le Meunier-FitzHugh, L. C. Le Meunier-FitzHugh, S. Misra, E. Mogaji, S. K. Sharma, J. B. Singh, V. Raghavan, R. Raman, N. P. Rana, S. Samothrakis, J. Spencer, K. Tamilmani, A. Tubadji, P. Walton, and M. D. Williams, "Artificial intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy," *International Journal of Information Management*, vol. 57, p. 101994, 2021.

[13] L. Cao, "AI in finance: Challenges, techniques, and opportunities," *ACM Computing Surveys*, vol. 55, no. 3, pp. 1–38, 2023.

[14] T. M. e. Silva Filho, H. Song, M. Perello-Nieto, R. Santos-Rodriguez, M. Kull, and P. A. Flach, "Classifier calibration: A survey on how to assess and improve predicted class probabilities," *Machine Learning*, vol. 112, no. 9, pp. 3211–3260, 2023.

[15] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, p. 6, 2020.

[16] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[17] G. Vilone and L. Longo, "Notions of explainability and evaluation approaches for explainable artificial intelligence," *Information Fusion*, vol. 76, pp. 89–106, 2021.

[18] C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong, "Interpretable machine learning: Fundamental principles and 10 grand challenges," *Statistics Surveys*, vol. 16, pp. 1–85, 2022.

[19] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable AI for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, 2020.

[20] S. Minaee, N. Kalchbrenner, E. Cambria, N. Nikzad, M. Chenaghlu, and J. Gao, "Deep learning–based text classification: A comprehensive review," *ACM Computing Surveys*, vol. 54, no. 3, pp. 1–40, 2021.

[21] E. Hüllermeier and W. Waegeman, "Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods," *Machine Learning*, vol. 110, pp. 457–506, 2021.

[22] Consumer Financial Protection Bureau, "Consumer complaint database," U.S. Consumer Financial Protection Bureau, 2025.

[23] Consumer Financial Protection Bureau, "2024 consumer response annual report," U.S. Consumer Financial Protection Bureau, 2025.