Financial Technology and Innovation

♦ ISSN: 2836-5372 Vol: 05, No. 02, 2025 ♦ PP. 01–09

Article DOI: https://doi.org/10.54216/FinTech-I.050201

FULL LENGTH ARTICLE

Remedy-Aware Financial Innovation: A Hierarchical RegTech Framework for Consumer Complaint Resolution

Laith Farhan[1,*] **Raad S. Alhumaima**[2]

[1] School of Engineering, Manchester Metropolitan University, Manchester, M1, UK

[2] Brunel University, Uxbridge UB8 3PH, UK

Emails: l.al-bayati@mmu.ac.uk · 1234914@alumni.brunel.ac.uk

Received: March 05, 2025 Revised: May 06, 2025 Accepted: August 04, 2025 ⋆ Corresponding author

ABSTRACT

Financial complaint systems keep track of instances of service failure as well as of the remedy chosen by the responding institution. The crucial question for operational supervision isn't just one of categorization of complaints, however; compensation is scarce, expensive, and fits into a larger framework of explanation vs. relief. In this paper, an approach for the development of a regulatory technology framework for remedies is proposed, which conceptualises the resolution of complaints as a hierarchy. The first stage determines if a case will be decided on relief or explanation, and the second stage distinguishes monetary from non-monetary relief. This is done by using sparse and regularized models, which are calibrated on the following validation period, and tested on an untouched chronological holdout, combining structured intake attributes with complaint narratives. In the empirical application, 268,570 complaints were received by California in 2024. Monetary relief makes up 1.04% of independent test period, and accuracy is not the best criterion. The hierarchical model yields a monetary-relief precision-recall area of 0.335, whereas the flat text-tabular model and a flat model with the event prevalence yield 0.314 and 0.010, respectively. At the 2 percent review budget, it is able to identify 58.9 percent of monetary-relief cases and has a lift of 29.4 times over a random review. The flat fusion model is slightly better for overall three-class classification, demonstrating the benefit of a hierarchy of remedies when institutional capacity is focused on rare, consequential outcomes, rather than labelling. Results provide a practical design of human-supervised complaint triage while retaining calibration, interpretability, chronological validation, and clear usage limitations for automated complaint analysis.

Keywords: Financial innovation ▪ Regulatory technology ▪ Consumer complaints ▪ Hierarchical classification ▪ Monetary relief ▪ Text analytics ▪ Responsible artificial intelligence

1. INTRODUCTION

Digital financial services have helped consumers see, reach and move with greater speed, range and diversity in their transactions, but they have also made more and more complex transactions potentially problematic. This makes financial institutions and supervisory bodies' information-processing challenge complex: complaints are received from a variety of sources, they address a variety of issues and concerns, and the eventual remedy is quite different. Typically, the information on these cases is considered to be "administratively the same" and put into standard case queues.These cases are usually given into the general case queue without manual review until they are found to be not administratively the same. If a subset of cases is correlated with monetary compensation, or other consequential measures of relief, that design may be inefficient.

Data from complaints is more than just for report. Complaint

data can help identify service issues, product weaknesses and institutional risk concentrations that can be detected earlier. Indeed, consumer-protection mechanisms and trust are strengthened by public disclosure [1], and the local number of complaints is a measure of the level of trust and effectiveness of these mechanisms [2]. However, predicting with such records must be done with caution. Complaint allegations are not a part of the evidence for determining actual facts; company response labels are not a measure of legal merit; and data-driven systems can be designed to replicate institutional patterns without accounting for why they are replicated. Any helpful RegTech system should be able to facilitate the allocation of reviews while not turning into an automated adjudicator.

There is a wealth of financial machine-learning research that has been dedicated to credit scoring and default prediction, lending, and fraud prediction [3–6]. Complaint analytics has been given less attention, and much past research has approached the problem by attempting classification of the product, binary outcome prediction, or general ranking of urgency [7–9] etc. These formulations have not yet exploited the semantic structures of the outcomes of resolution. Three labels are not separate nor unrelated: explanation, non-monetary remedy, and monetary remedy. Both the latter two show a wider relief state and the monetary relief is a less common conditional outcome of that state. In such a hierarchy, it may be better to allocate resources in such a way that missing out on scarce, high-consequence cases is more costly than reviewing additional cases; this can be improved by modeling this hierarchy.

This paper proposes a complaint triage framework for financial-sector innovation that focuses on remedies. It makes 4 contributions. First, it turns the problem of resolution prediction into a two-stage decomposition of probability instead of a single flat classification problem. Second, it combines structured intake data with complaint language, but excludes information generated once a company has handled the complaint. Third, there is a deployment under temporal change using a chronological training-validation-test design. Fourth, it assesses the system at specific review capacity levels and associates predictiveness with the number of cases an oversight team could reasonably review.

It is shown from the evidence that a useful boundary condition exists. The hierarchical model increases the accuracy of the rare monetary-remedy ranking and the low-budget capture, while it doesn't outperform the flat model in all metrics. Overall, conventional fusion model shows a higher accuracy score, as well as a higher score for balanced accuracy and macro F1. This is significant as the choice of the model should not precede the operational objective. In either case, the objective is broad resolution classification, the flat fusion would work best, and if the objective is to focus limited review capacity on possible monetary remedies, the hierarchical design would be most effective.

2. RELATED WORK AND RESEARCH GAP

2.1 Financial innovation and algorithmic decision support

Financial innovation increasingly combines alternative data, machine learning, and automated decision support. Digital footprints can enhance credit prediction [3], and FinTech lending has altered screening and access patterns [4, 10]. These developments create efficiency gains but also raise questions about unequal effects, transparency, and governance [5, 11, 12]. Broader reviews identify explainability, robustness, privacy, and regulatory alignment as central challenges for artificial intelligence in finance [13–15].

Explainable financial models have accordingly become an important research stream [16, 17]. Local and global explanation methods can reveal influential predictors [18], but post-hoc explanation is not a substitute for intelligible problem formulation or responsible deployment [19, 20]. In complaint triage, the model architecture itself should correspond to the decision structure. A hierarchy that separates relief from its form is more transparent than forcing three operationally related outcomes into a single undifferentiated choice.

2.2 Complaint analytics and regulatory technology

Consumer complaint research spans institutional behavior, regulatory disclosure, text classification, and operational prioritization. Hayes et al. [2] examine the relation between local trust and CFPB complaint activity. Dou and Roh [1] show that public complaint disclosure can affect market discipline. Within machine learning, Siering [7] develops explainable and fairness-aware RegTech for automated complaint monitoring, while Oyewola et al. [8] applies a two-stage residual convolutional network to CFPB complaint classification. Outside consumer finance, Vairetti et al. [9] combine deep language models and multicriteria decision-making to prioritize complaints.

Table 1 positions the present work against the closest streams. Three gaps remain. First, most studies do not model the nested structure of financial remedies. Second, broad predictive measures can obscure performance on rare monetary outcomes. Third, random partitions may overstate deployment performance where products, institutions, and complaint language change over time. The proposed design addresses these gaps through hierarchical probabilities, precision-recall and workload measures, and an out-of-time holdout.

3. RESEARCH DESIGN

3.1 Data boundary and analytical cohort

The analytical file is derived from the Consumer Financial Protection Bureau complaint database [22]. To ensure suitability for a 2025 volume, both evidence and data were frozen at 31 January 2025; no observation or reference dated after that boundary enters the analysis. The cohort contains California complaints received from 1 January through 31 December 2024. Restricting the analysis to one state holds constant a substantial portion of the legal and geographic environment while preserving variation across products, companies, issues, channels, and time.

Records are retained when the company response is one of three final categories: *closed with explanation*, *closed with non-monetary relief*, or *closed with monetary relief*. Untimely and unresolved responses are excluded because they do not represent the same terminal remedy choice. The resulting cohort contains 268,570 records. Complaint narratives are available for 30.0% of observations; the remaining records

Table 1. Selected related work and the remaining methodological gap

Study	Domain and data	Primary method	Main contribution	Limitation relative to the present study
Hayes et al. [2]	CFPB complaints and local trust	Econometric analysis	Connects complaint activity to social norms and consumer-protection effectiveness	Explains complaint incidence rather than remedy-aware operational triage
Siering [7]	CFPB regulatory monitoring	Explainable machine learning	Examines explainability and fairness in automated complaint supervision	Does not decompose explanation, non-monetary relief, and monetary relief as a hierarchy
Oyewola et al. [8]	CFPB complaint text	Two-stage residual 1D CNN	Improves automated classification from complaint narratives	Focuses on classification performance rather than chronological remedy capture under workload limits
Vairetti et al. [9]	Organizational complaints	Transformer models and multicriteria decision-making	Integrates automated labeling with urgency prioritization	Uses a different domain and expert-derived priority labels rather than observed financial remedies
Bussmann et al. [17]	Credit risk	Explainable machine learning	Demonstrates interpretable risk prediction in finance	Credit risk is structurally different from complaint resolution and company response behavior
Silva Filho et al. [21]	General classification	Calibration review	Consolidates methods for assessing and improving predicted probabilities	Provides methodology but no financial complaint application
Fuster et al. [5]	Mortgage credit markets	Machine learning and distributional analysis	Shows how predictive technologies may generate unequal market effects	Does not address post-transaction consumer-remedy operations
Černevičienė and Kabašinskas [14]	Finance literature	Systematic review	Maps the state of explainable AI in finance	Finds limited complaint-specific evidence and no capacity-aware remedy hierarchy
Present study	2024 financial complaints	Calibrated hierarchical text-tabular classification	Links nested remedy probabilities to chronological and budget-constrained evaluation	Limited to observed company responses in one state and year

The comparison emphasizes differences in decision target, validation design, and operational evaluation rather than claiming that the cited studies pursued identical objectives.

Table 2. Chronological sample design and outcome composition

Period	Total	Explanation	Non-monetary	Monetary
Train: Jan-Aug	156,747	75,045	79,216	2,486
Validation: Sep-Oct	53,737	28,235	24,902	600
Test: Nov-Dec	58,086	30,179	27,304	603
Full cohort	268,570	133,459	131,422	3,689

remain in the study because structured intake fields are available and narrative availability itself is observable at intake.

The split is strictly chronological. January-August records form the training period, September-October records form a calibration and model-selection period, and November-December records form the untouched test period. This design prevents future records from informing past predictions and provides a more credible approximation of operational deployment than random cross-validation.

3.2 Outcome hierarchy

Let $Y_i \in \{E, N, M\}$ denote explanation, non-monetary relief, and monetary relief for complaint i. A conventional model estimates all three probabilities in one softmax function. The proposed model instead represents the outcome through two conditional decisions. The first-stage probability is

$$r_i = P(Y_i \in \{N, M\} \mid \mathbf{x}_i), \tag{1}$$

which measures the probability of any relief. The second-stage probability is

$$m_i = P(Y_i = M \mid Y_i \in \{N, M\}, \mathbf{z}_i), \tag{2}$$

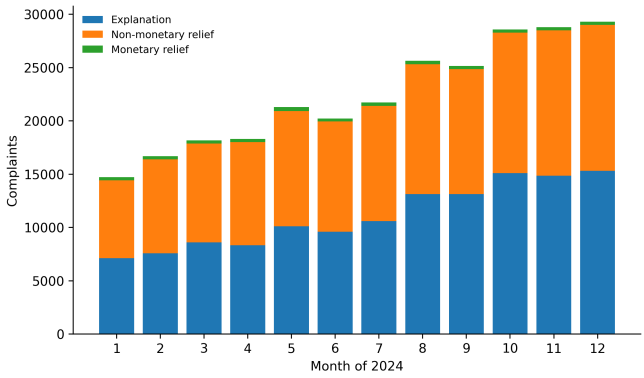
which measures the probability that relief is monetary. The final class probabilities are

$$P(Y_i = E) = 1 - r_i, \tag{3}$$
$$P(Y_i = N) = r_i(1 - m_i), \tag{4}$$
$$P(Y_i = M) = r_i m_i. \tag{5}$$

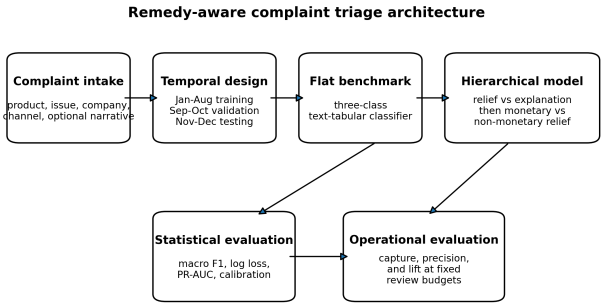
This decomposition respects the operational relationship between outcomes and isolates the severely imbalanced monetary decision within the subset where relief is plausible.

3.3 Predictors and leakage control

The structured feature set includes product, sub-product, issue, sub-issue, company, submission channel, consumer tags, narrative availability, narrative length, day of week, hour of receipt, and weekend status. Categories are learned from the training period only. Companies with fewer than 75 training records are pooled into an "other company" level, and infrequent one-hot levels are regularized through minimum-frequency encoding. Numeric fields are median-imputed and standardized.



(a) Monthly distribution of observed resolution outcomes.



(b) Leakage-safe chronological modeling and evaluation workflow.

Figure 1. Empirical setting and research workflow. The final two months are not used in model fitting, feature construction, or calibration.

Complaint narratives are transformed using term frequency-inverse document frequency with lowercase alphabetic tokens, English stop-word removal, sublinear term frequency, a minimum document frequency of ten, and a 6,000-term vocabulary learned exclusively from training records. This representation follows established evidence that sparse and deep text models offer complementary strengths for classification [23]. The design deliberately excludes response text, timeliness, consumer dispute status, and any variable created after company processing. These controls reduce target leakage and ensure that predictions could, in principle, be generated when a complaint enters the queue.

3.4 Models and probability calibration

Five specifications are compared:

1. a training-prior baseline;
2. a structured multiclass model;
3. a narrative-only multiclass model;
4. a flat fusion model combining structured and text features; and
5. the proposed hierarchical fusion model.

The predictive estimators are regularized linear classifiers optimized with stochastic gradient descent and log loss. Sparse linear models are selected because they scale to high-dimensional language features, permit direct coefficient inspection, and avoid presenting architectural complexity as innovation in itself. The monetary class receives additional weight during training. In the hierarchy, the relief stage uses fused structured-text features, while the conditional monetary stage uses structured fields. This arrangement was selected on the validation period because the text signal improved broad relief identification but did not consistently improve conditional monetary separation.

Stage probabilities are calibrated with Platt scaling on September-October observations. Calibration matters because risk scores are used for ranking and workload planning, not only hard labels [21]. The validation period is not reused in the final test metrics.

3.5 Evaluation criteria

Overall performance is reported through accuracy, balanced accuracy, macro F1, weighted F1, and multiclass log loss. Because monetary relief is rare, its precision-recall area (PR-AUC) is the primary rare-event criterion; reliance on accuracy or a single threshold metric would obscure minority-class performance [24]. Receiver operating characteristic area is included but interpreted cautiously because it can appear high under severe class imbalance. Relief PR-AUC evaluates the first-stage distinction.

Operational value is measured at review budgets $q \in \{0.5, 1, 2, 5, 10\}\%$. Complaints are ranked by predicted monetary probability. For the top q fraction, precision is the share actually closed with monetary relief, capture is the share of all monetary-relief cases found, and lift is

$$\text{Lift}(q) = \frac{\text{Precision}(q)}{P(Y = M)}. \tag{6}$$

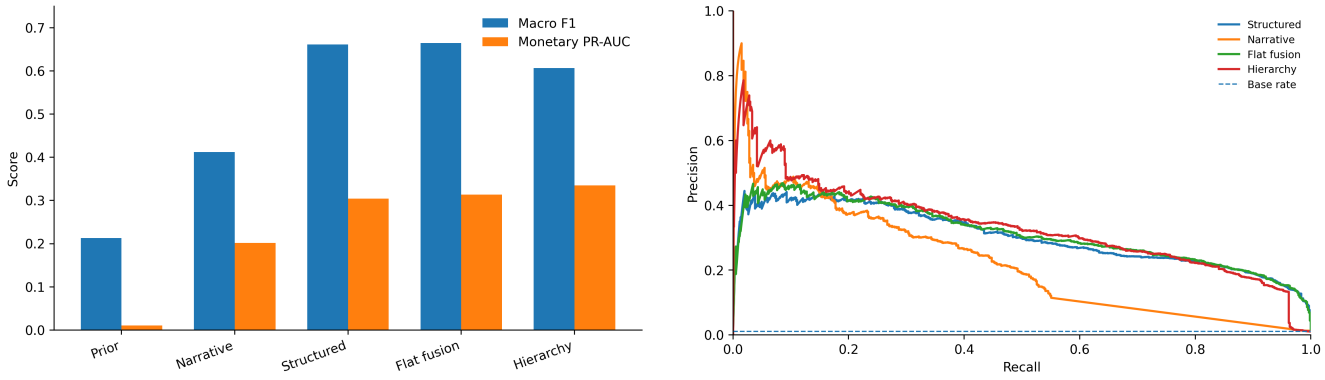
This formulation converts model quality into an interpretable staffing question: how many consequential cases can be identified when only a fixed portion of the queue can receive enhanced review?

4. EMPIRICAL RESULTS

4.1 Resolution patterns

The full cohort is dominated by credit-reporting complaints, which account for 227,036 observations. Their monetary-relief rate is only 0.043%, although non-monetary relief is common. By contrast, monetary outcomes are more prevalent for prepaid cards (25.1%), credit cards (16.3%), checking or savings accounts (16.2%), and money-transfer or virtual-currency services (9.1%). These differences establish why raw queue order is inefficient: complaint volume and monetary-remedy propensity are not aligned.

The independent test period contains 603 monetary-relief cases, a prevalence of 1.04%. A model that predicts only the most common class would therefore provide an apparently nontrivial accuracy while having no ability to locate monetary outcomes. Figure 2 compares the models using rare-event and general classification criteria.



(a) Out-of-time model comparison across key metrics.

(b) Precision-recall curves for monetary relief.

Figure 2. Predictive performance on complaints received in November-December 2024. PR-AUC is the central metric for the rare monetary outcome.

4.2 Model comparison

Table 3 shows a clear division between general classification and rare-event ranking. Flat fusion achieves the highest accuracy (0.803), balanced accuracy (0.705), and macro F1 (0.665). Structured features alone perform nearly as well, indicating that product, issue, company, channel, and intake timing contain most of the broad resolution signal. Narrative-only classification is substantially weaker because text is missing for roughly 70% of the cohort and because complaint language does not independently encode the institutional remedy process.

The hierarchical model produces the highest monetary PR-AUC, 0.335, representing more than thirty-two times the event prevalence. Its relief PR-AUC is also the highest at 0.848, and its log loss is the lowest at 0.463. The monetary ROC-AUC is lower than that of the flat model, illustrating why ROC-based conclusions alone would be incomplete. The hierarchy concentrates probability more effectively in the extreme upper tail, where review decisions occur, despite producing a weaker global ranking across all possible thresholds.

At the default maximum-probability threshold, the hierarchical model identifies explanation and non-monetary cases well: their F1 scores are 0.805 and 0.805, respectively. Monetary-relief precision is 0.461, but recall is 0.136 because the default decision rule is conservative under extreme imbalance. This is not the intended deployment threshold. The operational analysis instead ranks cases and applies a review budget, allowing decision makers to choose sensitivity according to available capacity.

4.3 Capacity-constrained monetary-remedy detection

Figure 3 and Table 4 present the central operational result. With capacity to review 0.5% of the test queue, the hierarchical model selects 291 complaints and finds 127 monetary-relief cases. Precision is 43.6%, capture is 21.1%, and lift is 42.0. At a 1% budget it captures 36.3% of monetary cases; at 2%, it captures 58.9% with 30.6% precision and 29.4-fold lift.

The hierarchy outperforms flat fusion at budgets from 0.5% through 2%, the region most relevant to highly constrained enhanced review. At 5% and 10%, flat fusion captures more monetary cases, reaching 90.5% and 99.2%, respectively. This crossover confirms that the hierarchy is not uniformly superior. It is a targeted innovation for concentrating the highest-risk tail, whereas the flat model provides broader coverage once capacity expands.

4.4 Interpretation of predictive associations

The coefficient analysis in Figure 4 is descriptive, not causal. Positive narrative terms include “fees,” “refund,” “charged,” and “fee,” all of which are semantically consistent with direct financial loss and requested reimbursement. Terms such as “debt,” “notice,” and “number” are negatively associated with monetary relief after controlling for structured fields, reflecting the large volume of credit-reporting and collection disputes that more often end in explanation or non-monetary correction.

Company and issue coefficients also vary substantially. These estimates must not be interpreted as institutional wrongdoing or intrinsic company quality. They combine complaint mix, product portfolio, response policy, reporting behavior, case selection, and unobserved severity. Their value is operational: historical patterns help locate cases whose observed intake profile resembles complaints that previously ended with monetary compensation. A responsible interface should therefore show influential factors and uncertainty while requiring human review before escalation.

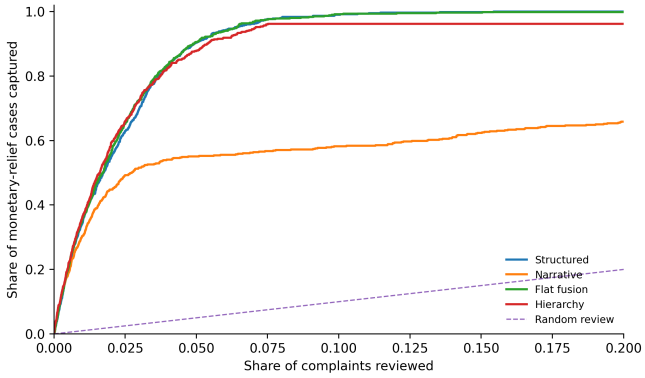
4.5 Calibration and subgroup robustness

Calibration is most consequential in the low-probability region because the event is rare. Figure 5a compares predicted and observed monetary-relief frequencies across quantile bins. The Platt-calibrated hierarchy tracks the empirical rate reasonably in lower and middle bins, while the highest-risk bin remains more variable because it concentrates a small number of heterogeneous cases. The model should thus support relative prioritization and workload planning rather than be treated as an exact probability of entitlement.

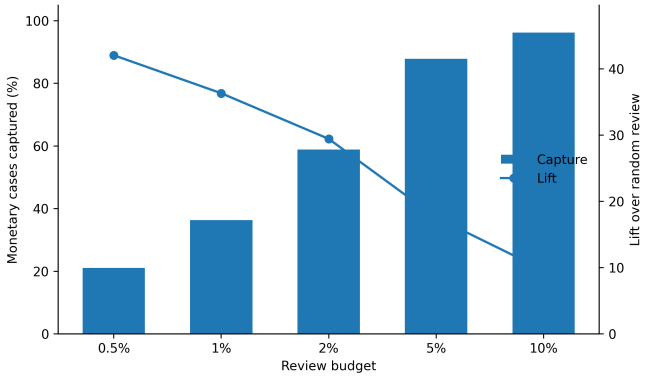
Subgroup results reveal substantial heterogeneity. Monetary PR-AUC is highest for prepaid cards (0.523) and credit cards (0.412), lower for checking or savings accounts (0.293), and very low for debt collection, where monetary outcomes are scarce. Performance is not directly comparable across groups with sharply different prevalence. Narrative-present complaints have a monetary rate of 2.24%, compared with 0.61% when no narrative is available; however, the hierar-

Table 3. Out-of-time predictive performance

Model	Accuracy	Bal. acc.	Macro F1	Weighted F1	Log loss	Monetary PR-AUC	Monetary ROC-AUC	Relief PR-AUC
Prior baseline	0.470	0.333	0.213	0.301	0.746	0.010	0.500	0.480
Narrative	0.527	0.459	0.412	0.453	0.768	0.202	0.803	0.524
Structured	0.802	0.691	0.661	0.802	0.497	0.304	0.982	0.835
Flat fusion	**0.803**	**0.705**	**0.665**	**0.803**	0.493	0.314	**0.982**	0.836
Hierarchical fusion	0.801	0.584	0.606	0.799	**0.463**	**0.335**	0.959	**0.848**



(a) Cumulative share of monetary remedies found as the reviewed queue expands.



(b) Precision and capture at operational review budgets.

Figure 3. Capacity-aware performance. The hierarchical model is strongest in the extreme upper tail; flat fusion overtakes it at broader review levels.

Table 4. Hierarchical model performance by review capacity

Budget	Reviewed	Precision	Capture	Lift
0.5%	291	0.436	0.211	42.0
1.0%	581	0.377	0.363	36.3
2.0%	1,162	0.306	0.589	29.4
5.0%	2,905	0.182	0.879	17.6
10.0%	5,809	0.100	0.962	9.6

chical PR-AUC remains meaningful in both groups. These findings support product-specific monitoring and threshold review rather than a single universal operating point.

5. DISCUSSION

5.1 Contribution to financial innovation research

The findings extend financial innovation research from origination decisions to post-transaction consumer operations. Credit-scoring studies show how digital information changes screening [3, 17]; the present evidence shows how complaint information can change the allocation of supervisory attention after a financial service has been delivered. The innovation is organizational as much as algorithmic. A remedy-aware queue converts static complaint repositories into a forward-looking decision-support layer while preserving human authority over investigation and resolution.

The hierarchy also demonstrates that model structure should reflect the institutional process. Monetary and non-monetary relief share a parent concept that is absent from a flat label representation. Encoding that relationship improves precision-recall performance in the extreme upper tail, where scarce review resources are deployed. At the same time, the flat model's stronger macro F1 cautions against assuming that a theoretically structured architecture will dominate every objective. Financial AI evaluation should report such trade-offs rather than selecting one favorable metric.

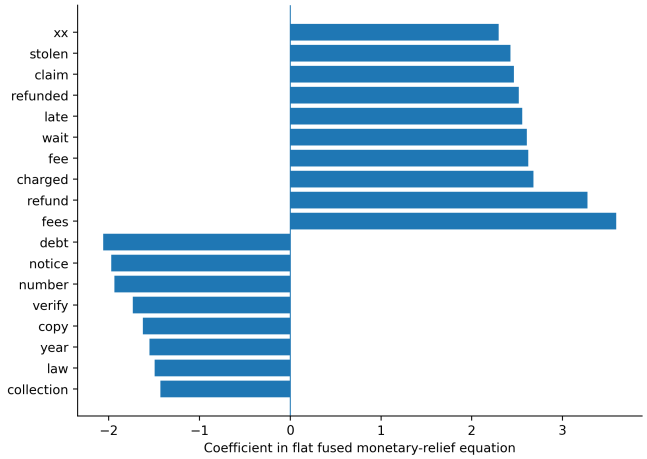
5.2 Operational implications

For a complaint office processing approximately 58,000 cases over two months, a 2% enhanced-review budget corresponds to 1,162 cases. The hierarchical model would have placed 355 of the 603 eventual monetary-remedy cases in that subset. This does not mean that those cases should receive money or that the remaining cases lack merit. It means that the queue can be organized so that analysts encounter a larger share of historically consequential cases earlier.

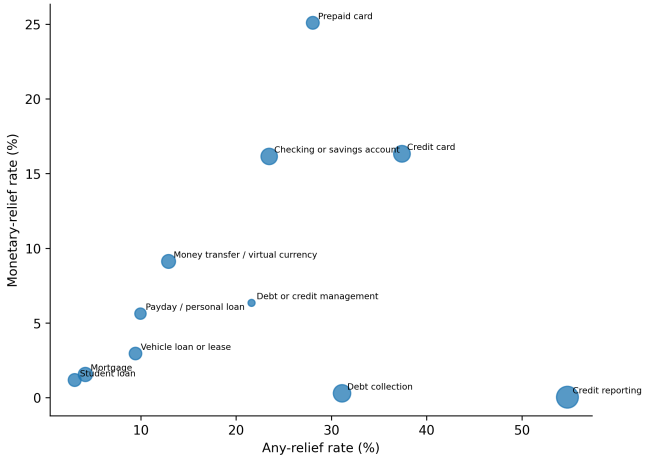
A practical implementation could combine three layers. The first is routine processing for the full queue. The second is an enhanced-review queue defined by a transparent capacity threshold. The third is portfolio monitoring that tracks shifts in product, issue, and institution distributions. Thresholds should be recalibrated periodically because complaint composition and company behavior can change. Model cards should document the data window, excluded variables, performance by product, and intended use.

5.3 Responsible deployment

Automated complaint triage is a high-accountability application. Explainability should therefore be treated as a governance mechanism, not a decorative chart [20, 25]. Each ranked case should retain its original information, display the strongest structured and textual signals, and allow reviewers to override the score. Firms or regulators should audit whether error rates differ across protected or vulnerable groups when legally and ethically appropriate data are available. The present file does not contain the demographic information needed for such an audit, so fairness cannot be

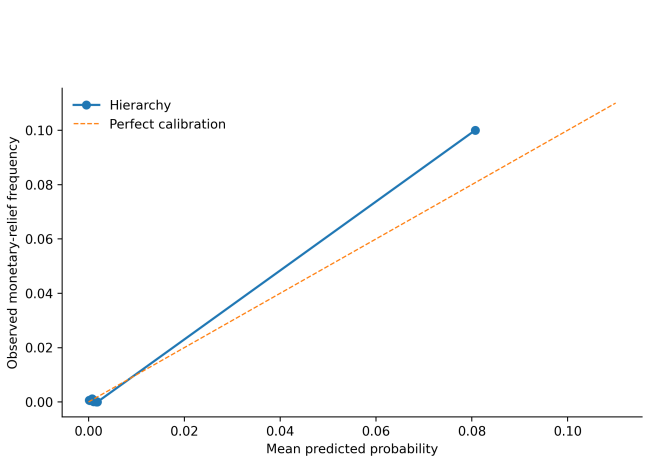


(a) Narrative terms most strongly associated with monetary relief in the flat fusion model.

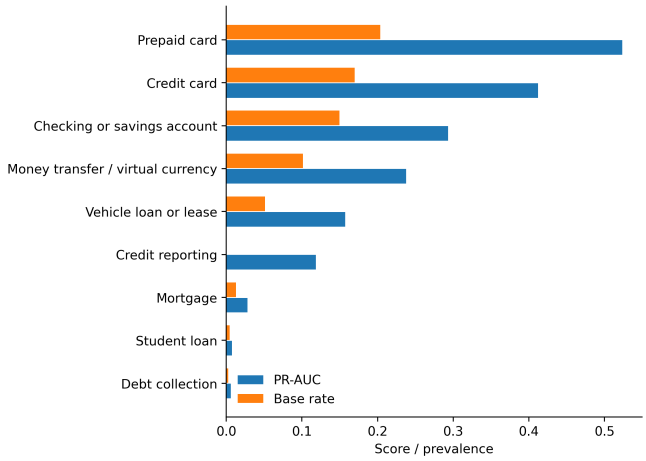


(b) Complaint volume and observed remedy rates by product.

Figure 4. Interpretability and product heterogeneity. Coefficients represent conditional predictive associations, not causal effects or findings of misconduct.



(a) Quantile calibration of hierarchical monetary-remedy probabilities.



(b) Monetary PR-AUC and prevalence across major product groups.

Figure 5. Probability reliability and subgroup variation on the chronological holdout.

inferred from aggregate product results.

The system should never be used to reject, close, or reduce the priority of an individual complaint without substantive review. Its defensible role is one-sided assistance: surfacing records for earlier attention. This asymmetry reduces the risk that imperfect predictions deny consumers access to review. Uncertainty, selective narrative submission, and institutional response policy must remain visible in any deployment decision [26, 27].

6. LIMITATIONS

The conclusions are limited by a number of restrictions. First, the result is not a stand-alone conclusion of consumer injury or legal rights, but the company's own response. Second, the analysis is limited to California and to just one calendar year, and extensions to other jurisdictions and years must be validated again. Thirdly, narratives are selectively published and missing for most complaints, which results in non-random missingness. Fourth, company identity is predictive, though it can contain effects of the portfolio mix, influence of internal policy and selection effect of complaints, which cannot be disentangled from this record. Fifth, the chronological test is not a multi-year regime change test, it is a two-month test. Lastly, there are no protected-class attributes, which means the study will not be able to conduct a full fairness analysis.

These restrictions do not mean that the result of the operation is incorrect, but that it gives a different interpretation. The framework anticipates observed remedies through an administrative process. It does not measure causation of harm, compensation for the victim, consumer welfare. Future research should be done to test for transportability across states and years, evaluate drift-aware recalibration, integrate process-time information available at legitimate decision points, and prospective human-in-the-loop trials.

7. CONCLUSION

This paper kind of develops a remedy aware RegTech framework for dealing with financial complaints, not just the generic stuff. It does that by splitting the problem into relief detection, and then conditional prediction of a monetary remedy, so the model kind of respects the statistical layout while still matching what the institution actually means by each outcome. On a clean November–December 2024 test slice, the hierarchy gets a monetary PR-AUC around 0.335,

and it also manages to catch about 58.9

So the takeaway is not really that one architecture is always best, but rather hierarchical design seems most useful when the real operational goal is early discovery of rare and consequential remedies when capacity is severely constrained. If you pair it with transparent features, periodic calibration, careful subgroup monitoring, and a mandatory human review step, the system becomes a more usable financial innovation for turning complaint data into accountable supervisory attention.

DATA AVAILABILITY

The exact analytical file used in the study is included in the accompanying reproducibility package as a compressed CSV file. It contains only the fields required to reproduce the results and is derived from the 2024 Consumer Complaint Database extract.

CODE AVAILABILITY

Executable Python code for preprocessing, chronological splitting, model estimation, calibration, evaluation, tables, and all PNG figures is included in the package.

ETHICS STATEMENT

The analysis uses de-identified administrative records. No attempt was made to identify complainants. The proposed system is limited to human-supervised prioritization and is not intended to determine complaint validity or consumer entitlement.

REFERENCES

[1] Y. Dou and Y. Roh, "Public disclosure and consumer financial protection," *Journal of Financial and Quantitative Analysis*, vol. 59, no. 5, pp. 2164–2198, 2024.

[2] R. M. Hayes, F. Jiang, and Y. Pan, "Voice of the customers: Local trust culture and consumer complaints to the CFPB," *Journal of Accounting Research*, vol. 59, no. 3, pp. 1077–1121, 2021.

[3] 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.

[4] 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.

[5] 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.

[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] M. Siering, "Explainability and fairness of RegTech for regulatory enforcement: Automated monitoring of consumer complaints," *Decision Support Systems*, vol. 158, p. 113782, 2022.

[8] D. O. Oyewola, T. O. Omotehinwa, and E. G. Dada, "Consumer complaints of consumer financial protection bureau via two-stage residual one-dimensional convolutional neural network (TSR1DCNN)," *Data and Information Management*, vol. 7, no. 4, p. 100046, 2023.

[9] C. Vairetti, I. Aránguiz, S. Maldonado, J. P. Karmy, and A. Leal, "Analytics-driven complaint prioritisation via deep learning and multicriteria decision-making," *European Journal of Operational Research*, vol. 312, no. 3, pp. 1108–1118, 2024.

[10] 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.

[11] 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.

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

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

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

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, 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.

[21] T. M. 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.

[22] Consumer Financial Protection Bureau, "Consumer complaint database: 2024 extract," Data set, 2024. [Online]. Available: https://www.consumerfinance.gov/data-research/consumer-complaints/

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.