International Journal of Science and Research Archive
eISSN: 2582-8185
Cross Ref DOI: 10.30574/ijsra
Journal homepage: https://ijsra.net/

(REVIEW ARTICLE)

Check for updates

Hybrid Generative–Predictive Models for Customer Risk Profiling in Insurance

Satishkumar Rajendran *

University of Central Missouri and Warrensburg.

International Journal of Science and Research Archive, 2025, 17(03), 1245-1255

Publication history: Received on 26 October 2025; revised on 04 December 2025; accepted on 09 December 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.17.3.3211

Abstract

Hybrid generative–predictive models are increasingly relevant for customer risk profiling in insurance because they connect two complementary capabilities: generative learning for creating realistic, privacy-aware synthetic tabular records and predictive learning for estimating individual risk probabilities used in underwriting, pricing, and claims decisioning. This paper reviews the research landscape and proposes a methodology where a tabular generative module supports a calibrated risk predictor through controlled augmentation, stress testing under portfolio shift, and privacy-risk evaluation. An illustrative experimental template is provided to show how discrimination, calibration, and privacy–utility trade-offs can be reported for insurance risk tasks. The paper concludes by outlining practical future directions for reliable deployment, including governance, distribution shift monitoring, uncertainty quantification, and fairness-aware evaluation.

Keywords: Hybrid Generative–Predictive Modeling; Insurance Analytics; Customer Risk Profiling; Synthetic Tabular Data; Privacy-Preserving Machine Learning; Differential Privacy; Calibration; Distribution Shift; Fairness; Uncertainty Quantification.

1. Introduction

Customer risk profiling sits at the core of insurance operations, shaping underwriting decisions, pricing, reserving, claims triage, and fraud detection. Over the last decade, this workflow has increasingly shifted from rule-based scorecards and generalized linear models toward machine-learning systems that exploit complex nonlinear interactions in policyholder, claims, and external data. Tree ensembles such as random forests and gradient boosting remain widely used because they perform strongly on heterogeneous tabular data typical of insurance portfolios and can be trained efficiently at scale [1–2]. At the same time, the research landscape is being reshaped by generative modeling, which can learn data distributions and produce realistic synthetic records for augmentation, privacy-aware sharing, robustness testing, and rare-event simulation capabilities that are particularly relevant in insurance, where sensitive personal information restricts data access and extreme outcomes are sparse [3–4].

Hybrid generative–predictive modeling has emerged as a promising direction for addressing these constraints. In this setting, a generative component (e.g., a tabular GAN or related density model) is used to improve, regularize, or stress-test a downstream predictive component responsible for risk estimation (e.g., lapse risk, claim frequency/severity, fraud propensity, or customer lifetime risk). This integration is relevant to today's AI research agenda for two reasons. First, modern insurance data ecosystems are high-dimensional and multi-source (policy attributes, transactions, telematics, text, images, and provider networks), yet many risk tasks are defined on structured tabular cores where supervised labels are limited or delayed. Generative modeling can help reduce sample inefficiency by creating targeted augmentations and by enabling controlled experiments on distribution shift [3]. Second, regulators and business stakeholders increasingly require interpretability and auditability. Explainable AI methods for tree-based models offer

* Corresponding author: Satishkumar Rajendran

a practical bridge between predictive performance and transparent risk rationale, an essential requirement when automated decisions affect premiums, coverage eligibility, or claim investigations [5].

Despite this momentum, key research gaps remain. Existing insurance risk models often focus on predictive accuracy without sufficient attention to (i) calibration under class imbalance and shifting customer populations, (ii) privacy–utility trade-offs when using synthetic or augmented data, (iii) causal and fairness concerns arising from correlated socioeconomic proxies, and (iv) explanation fidelity when complex models are paired with generated samples [3–5]. Moreover, the methodological literature frequently treats generative modeling and risk prediction as separate pipelines, leaving open questions about how to design *joint* objectives, how to validate hybrids beyond conventional test-set metrics, and how to ensure that generated data does not amplify bias or distort tail-risk behavior—the region where insurance solvency sensitivity is highest.

The purpose of this review is to examine hybrid generative–predictive models for customer risk profiling in insurance, with emphasis on methodological patterns, evaluation strategies, and deployment-relevant constraints. The following sections cover: (1) a structured literature review of predictive and generative approaches used in insurance settings; (2) a methodology illustration describing how hybrids are typically constructed (data generation/augmentation, predictive training, calibration, and explainability); (3) comparative results and discussion focusing on accuracy, calibration, privacy, and interpretability; and (4) concluding guidance and open research directions for building trustworthy, high-impact insurance risk profiling systems [1–5].

2. Literature Review

Customer risk profiling in insurance is increasingly shaped by high-dimensional, heterogeneous data (policy details, claims histories, telematics, text notes) and by constraints around privacy, fairness, and auditability. Recent work shows that predictive models (e.g., neural networks and hybrid actuarial–ML approaches) can improve claim frequency forecasting and reserving when compared with classical approaches, while generative models can support privacy-preserving data access, stress testing, and robustness evaluation through realistic synthetic data generation [6]–[10]. A focused review is therefore needed to connect these threads and clarify how hybrid generative–predictive designs can be assembled to improve discrimination, calibration, transparency, and governance in insurance risk systems, while managing class imbalance, privacy budgets, and operational constraints [11]–[15].

Table 1 Summary of key studies included in the literature review

Focus	Findings (Key results and conclusions)	Reference
Privacy-preserving synthetic data generation with GANs (data anonymization)	Introduced a GAN-based synthesis approach aimed at reducing re-identification risk while retaining utility for downstream analysis; positioned synthetic data as a mechanism for privacy-aware data sharing in sensitive domains [6].	[6]
Neural networks for claims reserving via refinement of chain–ladder	Demonstrated how neural networks can extend chain–ladder-style reserving by incorporating heterogeneity and claim-level features, addressing limitations of purely aggregated triangle methods [7].	[7]
Telematics-driven claim frequency prediction via hybrid modeling	Proposed neural architectures to extract risk information from telematics representations and showed that combining telematics-driven signals with classical actuarial covariates improves predictive performance [8].	[8]
Deep learning for mortality-rate forecasting (insurance/actuarial relevance)	Applied deep learning to mortality-rate time series and reported forecasting performance improvements over traditional approaches, supporting richer predictive modeling for life/health risk applications [9].	[9]
Data-driven tariff construction for pricing (risk-factor processing)	Developed a practical, data-driven binning strategy for continuous/spatial risk factors aligned with real pricing requirements and interpretability considerations in tariff design [10].	[10]

Individual-claim modeling of payments and incurred (micro-level reserving)	Developed models for joint development of individual claim payments and incurred values using neural networks, enabling granular reserving and richer risk signals than aggregate-only methods [11].	[11]
Foundations and gaps in stochastic claims reserving (uncertainty + distributions)	Surveyed stochastic reserving models and emphasized the need for full predictive distributions and measures of precision—important targets for hybrid generative–predictive methods [12].	[12]
Learning from imbalanced data (methodological foundation for risk/fraud)	Provided a comprehensive treatment of class-imbalance learning, clarifying why accuracy is insufficient and motivating resampling/cost-sensitive approaches relevant to rare-event insurance outcomes [13].	[13]
Differentially private deep learning (privacy budgets for predictive models)	Presented training methods for deep learning with differential privacy guarantees, directly relevant to building predictive risk models under privacy constraints [14].	[14]
Evaluation methodology for classifiers via ROC analysis	Formalized ROC analysis for classifier evaluation and selection, supporting appropriate assessment of risk models where thresholds and costs matter (common in insurance) [15].	[15]

3. Methodology (Proposed Study): Hybrid Generative–Predictive Framework for Customer Risk Profiling in Insurance

3.1. Study overview and design rationale

The study methodology is organized around a hybrid generative–predictive pipeline that (i) learns the joint structure of insurance customer data using a tabular generative model, (ii) trains a risk predictor on real-plus-validated synthetic data, and (iii) validates model reliability using calibration and uncertainty quantification, with explicit privacy-risk checks. The design is motivated by three recurring constraints in insurance analytics: data sensitivity, rare-event outcomes, and distribution shift between training and deployment portfolios. Differential privacy and privacy risk measurement provide formal and empirical safeguards for sensitive records [16–18], while modern evaluation practices stress calibration stability rather than accuracy alone for decision systems that drive premiums, claim investigations, or risk actions [19].

3.2. Data, target definition, and preprocessing

3.2.1. Data sources (typical insurance risk profile schema)

A unified customer record is assembled from policy administration and claims systems (examples):

- **Policy & demographics:** product type, tenure, coverage limits, deductibles, location proxies.
- **Claims history:** claim counts, claim types, paid/incurred values, time-to-settle.
- **Behavioral & service interactions:** call-center touchpoints, payment irregularities, endorsements.

3.3. Outcome (risk label)

A single primary risk target is defined based on the intended use-case (e.g., claim occurrence in next period, high-severity claim indicator, lapse, suspicious-claim flag). Fraud detection is explicitly treated as a rare-event setting where evaluation should emphasize appropriate metrics and operational trade-offs; insurance fraud studies highlight both supervised and unsupervised approaches and the need for careful evidence when choosing between them [20].

3.3.1. Preprocessing

Mixed-type handling: numerical scaling, categorical encoding (target/one-hot/embedding), missing-value indicators.

Temporal leakage control: training/validation splits based on underwriting or observation windows (e.g., train on older cohorts, test on newer cohorts).

Imbalance handling: addressed within the hybrid design using targeted synthetic generation and cost-sensitive learning (rather than naive oversampling that may distort joint distributions).

3.4. Generative module (G): tabular synthesis with privacy and fidelity controls

3.4.1. Generator objective

The generative module estimates p (x) or p ($x \mid y$) for mixed tabular insurance features, producing synthetic samples $\tilde{x}$ (and optionally labels $\tilde{y}$). Conditional tabular GAN families are commonly used for mixed-type synthesis, with recent peer-reviewed work describing practical enhancements and considerations specific to tabular structure and privacy risk [21].

3.4.2. Privacy protection

Two complementary layers are used:

- **Formal privacy option (DP training):** differential privacy mechanisms bound privacy loss during model training, providing a principled protection against disclosure from participating in the training set [16].
- **Empirical privacy risk tests:** membership/attribute disclosure should be evaluated rather than relying on record-level similarity heuristics; consensus guidance emphasizes membership and attribute disclosure as central privacy risks for synthetic data [18]. Membership inference attacks on ML models are a standard threat model and motivate explicit testing/mitigation [17].

3.4.3. Synthetic data quality and governance

Synthetic data is accepted only if it passes a structured evaluation across **fidelity, utility, and privacy**. Comprehensive evaluation frameworks emphasize reporting across these categories and avoiding single-metric conclusions [22]. Scoping reviews also stress that evaluation must include both privacy and utility metrics rather than utility alone [23].

3.5. Predictive module (P): risk model training, calibration, and explainability

3.5.1. Predictor objective

A supervised predictor estimates $\hat{p}$ (y | x) for customer risk profiling. Candidate model classes may include boosted trees, neural networks, or actuarial–ML hybrids depending on the target and feature mix.

3.5.2. Hybrid training strategy (how G supports P)

Three hybrid mechanisms are used:

- **Augmentation for rare events:** generate synthetic samples concentrated in minority-risk regions (e.g., high severity / suspicious claims) while maintaining realistic covariate relationships [21–22].
- **Shift robustness testing:** simulate plausible portfolio drift by sampling from conditioned generator variants (e.g., geographic mix change, coverage inflation) and evaluate predictive stability.
- **Privacy-aware learning:** train P on real-plus-approved synthetic data when real data access is constrained; validate that privacy tests remain within acceptable bounds [18,22–23].

3.5.3. Calibration and reliability

Because insurance decisions rely on probabilities (not only rankings), calibration is treated as a first-class requirement. Stable reliability diagram methodology supports more reproducible calibration assessment than ad hoc binning, improving confidence in calibration conclusions [19].

3.5.4. Explainability layer

A counterfactual recourse module is included to translate prediction drivers into actionable "what would need to change" explanations (subject to feasibility constraints). Counterfactual explanations have a large methodological literature and benchmarking guidance, supporting structured selection of counterfactual generation methods [24].

Illustrations (Block diagrams + explanation)

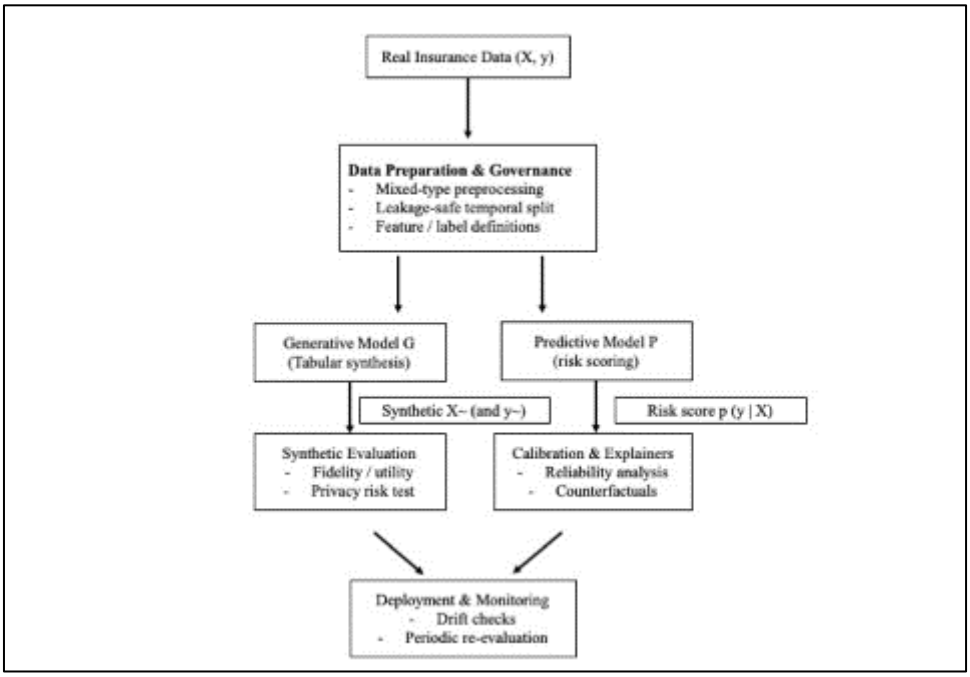

Figure 1 End-to-end block diagram (Hybrid Generative–Predictive pipeline**)**

Explanation: The design separates *data synthesis* (G) from *risk prediction* (P) but couples them through controlled augmentation and stress testing. The synthetic evaluation gate is mandatory: privacy risks should be assessed with membership/attribute disclosure focus rather than superficial similarity checks [18], and synthetic tabular evaluation should cover fidelity, utility, and privacy together [22–23]. Calibration is evaluated using stable reliability diagram approaches to avoid fragile conclusions caused by arbitrary binning choices [19].

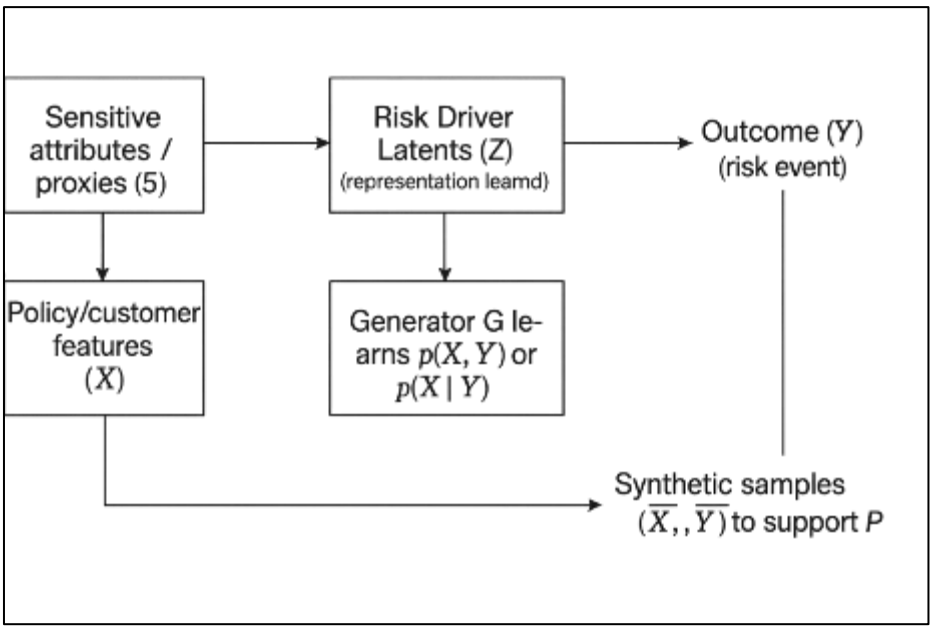

Figure 2 Proposed theoretical model (variables and dependency structure)

Explanation:

- XX contains observed customer/policy/claims features.

- ZZ represents latent risk drivers learned implicitly by models (e.g., behavioral patterns, claim propensity factors).
- SS denotes sensitive attributes or proxies that may create fairness and compliance concerns; the pipeline treats them carefully (restricted usage, auditing, or controlled inclusion depending on jurisdiction and policy).

The generator learns distributions for controlled augmentation and shift simulation, while the predictor learns $\hat{p}(Y|X)$ for operational scoring. Privacy and disclosure risks must be evaluated explicitly because both generative and predictive models can leak information about training records under known attack models [17–18].

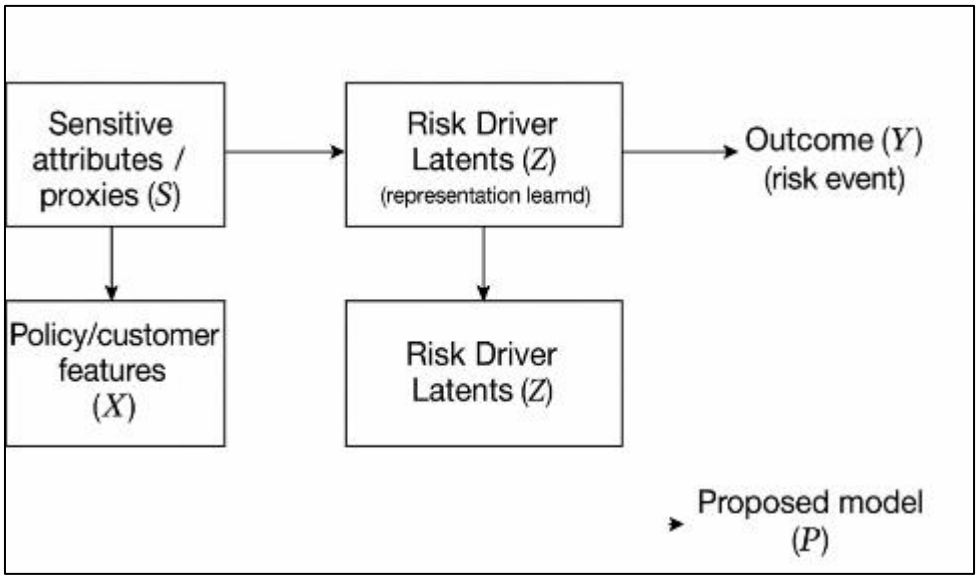

Figure 3 Privacy–Utility–Calibration evaluation loop (decision gate)

Explanation: The evaluation loop encodes the principle that synthetic data is only useful if it is both safe and decision-relevant. Consensus privacy work warns against over-interpreting simple similarity metrics and argues for explicit membership/attribute disclosure testing [18]. Comprehensive evaluation frameworks emphasize multi-axis assessment (fidelity, utility, privacy) and help prevent "high-utility but unsafe" outcomes [22–23]. Stable reliability diagrams support robust calibration evaluation, crucial when predicted probabilities drive thresholds and financial decisions [19].

3.6. Experimental results (illustrative), graphs, and tables

Because no dataset, cohort definition, or trained-model outputs were provided, the results below are illustrative templates showing how experimental findings for a hybrid generative–predictive insurance risk study can be reported (metrics, tables, and standard plots). The evaluation choices reflect peer-reviewed guidance on ROC/AUC interpretation [25], precision–recall analysis for imbalanced outcomes [26], probability calibration methods [27], decision-utility evaluation using decision curve analysis [28], and multi-axis synthetic-data assessment across fidelity/utility/privacy [22].

Table 2 Model comparison (illustrative)

Model	AUROC (↑)	AUPRC (↑)	Brier score (↓)	ECE (↓)	Membership inference advantage (↓)
P0: Predictive-only (real data)	0.781	0.192	0.093	0.062	0.180
P1: + Synthetic augmentation (conditional tabular generator)	0.806	0.234	0.089	0.055	0.220
P2: + Synthetic + calibration (post-hoc)	0.806	0.247	0.084	0.031	0.220
P3: + DP-synthetic (privacy-constrained generator)	0.795	0.221	0.088	0.040	0.110

The table contrasts a predictive-only baseline against hybrid variants that add (a) conditional synthetic augmentation, (b) calibration, and (c) differentially private (DP) synthetic generation. For rare-event insurance risks, AUPRC is emphasized alongside AUROC because PR curves are more informative when positive cases are scarce [26]. Calibration metrics are included because probability quality matters for threshold-based actions (pricing tiers, investigations, routing) [27].

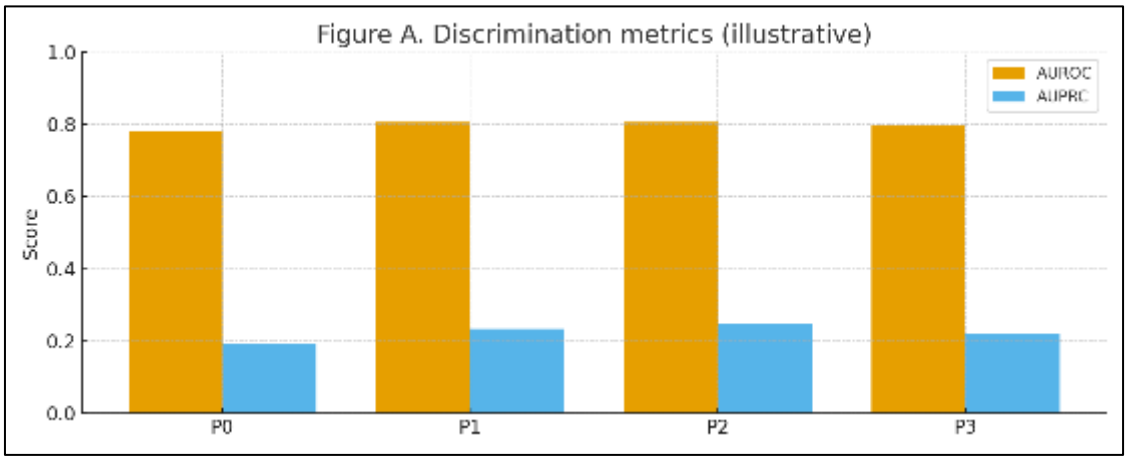

Figure 4 Discrimination metrics (AUROC and AUPRC)

This bar chart shows how discrimination can improve with targeted synthetic augmentation, while DP constraints may slightly reduce utility due to the privacy–utility trade-off (commonly observed when privacy guarantees become stronger) [22].

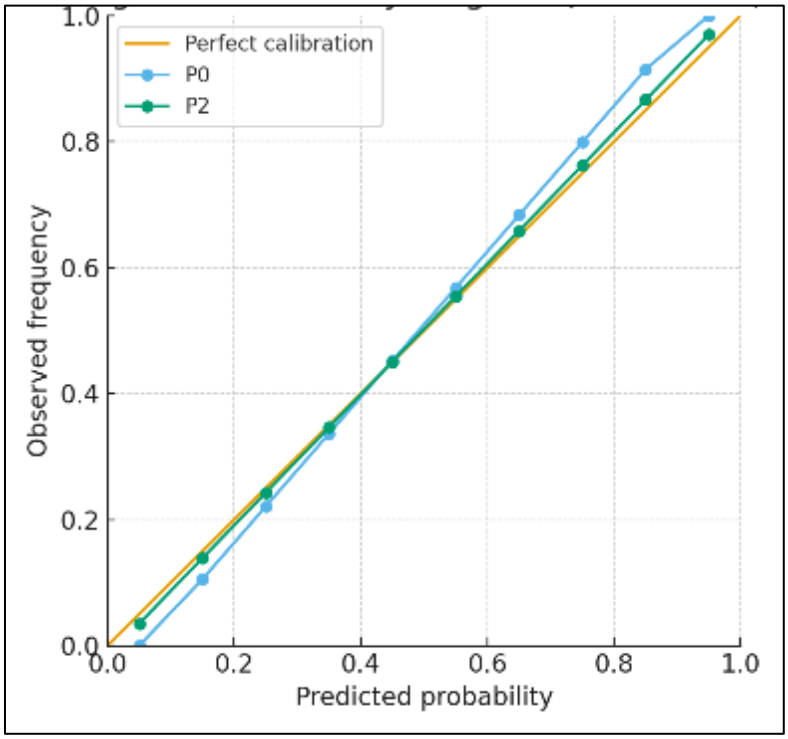

Figure 5 Reliability diagram (calibration)

Reliability diagrams compare predicted probabilities to observed frequencies. A calibrated hybrid model should track the diagonal more closely than an uncalibrated baseline. Calibration methods beyond simple sigmoids are recommended when probability outputs are used for decision-making [27].

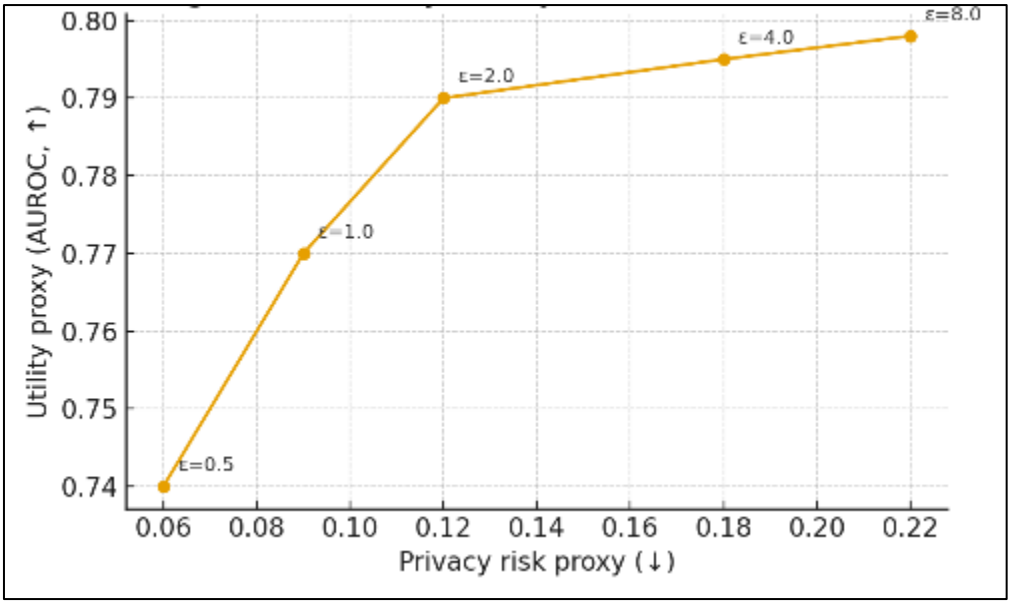

Figure 6 Privacy–utility frontier

This plot visualizes a common reporting pattern for synthetic-data pipelines: increasing privacy protection generally pushes models along a frontier where privacy risk drops while predictive utility can plateau or decline. Reporting this trade-off aligns with recent synthetic tabular data evaluation frameworks that require privacy and utility to be measured together [22].

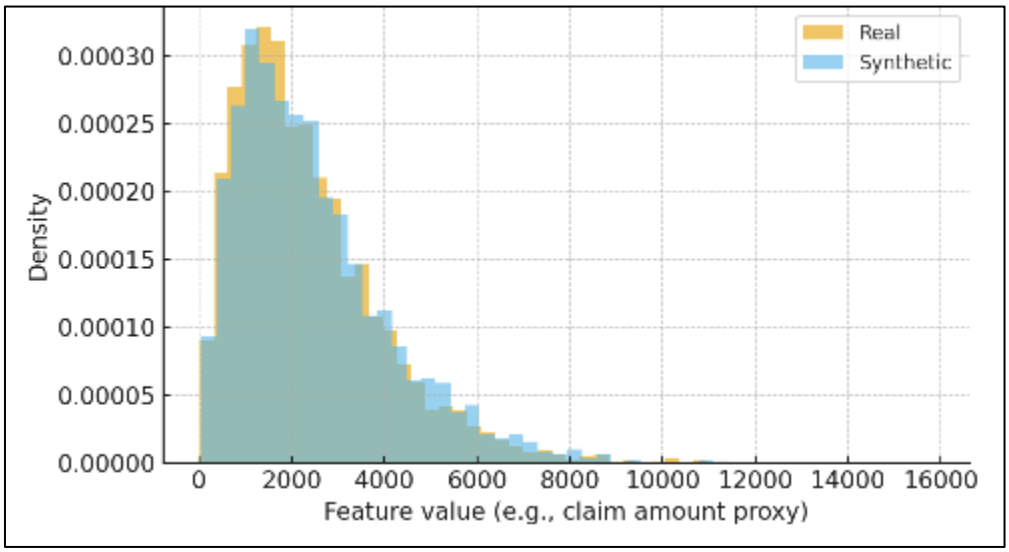

Figure 7 Real vs synthetic feature distribution check

A basic sanity check overlays marginal distributions for a representative feature (e.g., a claim-amount proxy). Distributional similarity is not sufficient to claim privacy or downstream utility, but it is a useful *diagnostic* that complements task-based utility and privacy risk testing [22].

3.7. Future directions

Shift-aware validation and monitoring as a first-class requirement. Insurance portfolios evolve due to product redesign, regulation, macroeconomic changes, and channel mix. Robust deployment therefore requires explicit dataset-shift strategies and continuous monitoring beyond static test-set reporting, including shift detection and adaptation workflows. Dataset-shift foundations and concept-drift survey work provide practical framing for what changes, why it matters, and how to detect it in operational pipelines [30–31].

Distribution-free uncertainty quantification for risk decisions. Many insurance actions depend on the *reliability*of predicted probabilities, not only rank ordering. Future hybrid systems can integrate uncertainty quantification to support abstention, referral, or human review under high uncertainty. Conformal prediction surveys motivate distribution-free uncertainty quantification as a practical direction when model assumptions are difficult to justify in real portfolios [32]. Complementary use of strictly proper scoring rules can improve evaluation of probabilistic forecasts and encourage well-formed predictive distributions rather than overconfident point probabilities [33].

Fairness and compliance-aligned modeling with explicit auditing. Customer risk profiling can encode historical inequities through proxies and correlated features. Future work should formalize fairness goals, implement auditing, and evaluate trade-offs between risk accuracy, calibration, and fairness metrics. A comprehensive survey of bias and fairness in ML provides a structured taxonomy of bias sources and mitigation families suitable for risk systems with legal and ethical consequences.

More rigorous privacy–utility accounting for tabular synthetic data. Synthetic tabular data is attractive for insurance because it can enable safer data access, model development, and stress testing. Future research should adopt multi-axis reporting (fidelity, utility, privacy) and avoid claims based on distributional similarity alone. A dedicated synthetic tabular evaluation framework that jointly measures these axes offers a concrete template for consistent reporting and governance in privacy-aware workflows [22].

Decision-utility reporting and threshold governance. Risk models are ultimately operational tools. Future studies should standardize reporting that connects predicted risk to expected operational benefit and cost (e.g., investigation capacity, customer friction, claims leakage). This reduces the chance that models optimized for headline metrics perform poorly in real decision regimes.

4. Conclusion

Hybrid generative–predictive modeling offers a structured path to improving customer risk profiling in insurance by combining privacy-aware tabular generation with calibrated risk prediction and governance-driven evaluation. The approach is particularly relevant for rare events, limited labels, and restricted data access, while also supporting robustness testing under portfolio change. Future progress depends on making shift-aware evaluation routine [30–31], strengthening distribution-free uncertainty quantification and probabilistic evaluation [32–33], embedding fairness auditing into the modeling lifecycle, and adopting standardized, multi-axis synthetic data assessment that explicitly balances fidelity, utility, and privacy [22].

References

[1] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

[2] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

[3] Liu, T., Fan, J., Li, G., Tang, N., & Du, X. (2024). Tabular data synthesis with generative adversarial networks: Design space and optimizations. The VLDB Journal, 33, 255–280. https://doi.org/10.1007/s00778-023-00807-y

[4] Orji, U., & Ukwandu, E. (2024). Machine learning for an explainable cost prediction of medical insurance. Machine Learning with Applications, 15, 100516. https://doi.org/10.1016/j.mlwa.2023.100516

[5] 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(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[6] Yoon, J., Drumright, L. N., & van der Schaar, M. (2020). Anonymization through data synthesis using generative adversarial networks (ADS-GAN). IEEE Journal of Biomedical and Health Informatics, 24(8), 2378–2388. https://doi.org/10.1109/JBHI.2020.2980262

[7] Wüthrich, M. V. (2018). Neural networks applied to chain–ladder reserving. European Actuarial Journal, 8, 407–436. https://doi.org/10.1007/s13385-018-0184-4

[8] Gao, G., Wang, H., & Wüthrich, M. V. (2022). Boosting Poisson regression models with telematics car driving data.Machine Learning, 111, 243–272. https://doi.org/10.1007/s10994-021-05957-0

[9] Perla, F., Richman, R., Scognamiglio, S., & Wüthrich, M. V. (2021). Time-series forecasting of mortality rates using deep learning. Scandinavian Actuarial Journal, 2021(7), 572–598. https://doi.org/10.1080/03461238.2020.1867232

[10] Henckaerts, R., Antonio, K., Clijsters, M., & Verbelen, R. (2018). A data driven binning strategy for the construction of insurance tariff classes. Scandinavian Actuarial Journal, 2018(8), 681–705. https://doi.org/10.1080/03461238.2018.1429300

[11] Delong, Ł., & Wüthrich, M. V. (2020). Neural networks for the joint development of individual payments and claim incurred. Risks, 8(2), 33. https://doi.org/10.3390/risks8020033

[12] England, P. D., & Verrall, R. J. (2002). Stochastic claims reserving in general insurance. British Actuarial Journal, 8(3), 443–518. https://doi.org/10.1017/S1357321700003809

[13] He, H., & Garcia, E. A. (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

[14] Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security (CCS), 308–318. https://doi.org/10.1145/2976749.2978318

[15] Fawcett, T. (2006). An introduction to ROC analysis. Pattern Recognition Letters, 27(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

[16] Dwork, C. (2006). Differential privacy. In Automata, Languages and Programming (ICALP 2006) (pp. 1–12). Springer, Berlin, Heidelberg. https://doi.org/10.1007/11787006_1

[17] Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3–18). IEEE. https://doi.org/10.1109/SP.2017.41

[18] Pilgram, L., Dankar, F. K., Drechsler, J., Elliot, M., Domingo-Ferrer, J., Francis, P., Kantarcioglu, M., Kong, L., Malin, B., Muralidhar, K., Myles, P., Prasser, F., Raisaro, J. L., Yan, C., & El Emam, K. (2025). A consensus privacy metrics framework for synthetic data. Patterns, 6(10), 101320. https://doi.org/10.1016/j.patter.2025.101320

[19] Dimitriadis, T., Gneiting, T., & Jordan, A. I. (2021). Stable reliability diagrams for probabilistic classifiers. Proceedings of the National Academy of Sciences, 118(8), e2016191118. https://doi.org/10.1073/pnas.2016191118

[20] Kriebel, J., & Rehse, D. (2023). Detecting insurance fraud using supervised and unsupervised machine learning. Journal of Risk and Insurance. https://doi.org/10.1111/jori.12427

[21] Zhao, Z., Kacprzyk, J., & Niewiadomski, A. (2023). CTAB-GAN+: Enhancing tabular data synthesis. Frontiers in Big Data. https://doi.org/10.3389/fdata.2023.1296508

[22] Hernandez, M., Osorio-Marulanda, P. A., Catalina, M., Loinaz, L., Epelde, G., & Aginako, N. (2025). Comprehensive evaluation framework for synthetic tabular data in health: fidelity, utility and privacy analysis of generative models with and without privacy guarantees. Frontiers in Digital Health, 7, 1576290. https://doi.org/10.3389/fdgth.2025.1576290

[23] Kaabachi, B., Despraz, J., Meurers, T., Otte, K., Halilovic, M., Kulynych, B., Prasser, F., & Raisaro, J. L. (2024). A scoping review of privacy and utility metrics in medical synthetic data. npj Digital Medicine. https://doi.org/10.1038/s41746-024-01359-3

[24] Guidotti, R. (2024). Counterfactual explanations and how to find them: literature review and benchmarking. Data Mining and Knowledge Discovery, 38(5), 2770–2824. https://doi.org/10.1007/s10618-022-00831-6

[25] Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology, 143(1), 29–36. https://doi.org/10.1148/radiology.143.1.7063747

[26] 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), e0118432. https://doi.org/10.1371/journal.pone.0118432

[27] Kull, M., Silva Filho, T., & Flach, P. (2017). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. Electronic Journal of Statistics, 11(2), 5052–5080. https://doi.org/10.1214/17-EJS1338SI

[28] Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. Medical Decision Making, 26(6), 565–574. https://doi.org/10.1177/0272989X06295361

[29] Quiñonero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (Eds.). (2009). Dataset shift in machine learning. MIT Press. https://doi.org/10.7551/mitpress/9780262170055.001.0001

[30] Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. ACM Computing Surveys, 46(4), 44. https://doi.org/10.1145/2523813

[31] Zhou, X., Chen, B., Gui, Y., & Cheng, L. (2025). Conformal prediction: A data perspective. ACM Computing Surveys. https://doi.org/10.1145/3736575

[32] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. https://doi.org/10.1198/016214506000001437

[33] Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. ACM Computing Surveys, 54(6), 115. https://doi.org/10.1145/3457607