African Journal of Education Science and Technology (AJEST) Vol. 8 No. 4 (May, 2026)
University of Eldoret, Kenya, Mount Kenya University, Kenya, Chukwuemeka Odumegwu Ojukwu University, Nigeria, Kyambogo University, Uganda and University of Makeni, Sierra Leone

Explainable Artificial Intelligence (XAI) in Actuarial Pricing and Reserving: A Systematic Literature Review of Methods, Regulatory Compliance, and Predictive Trade-offs

Faridah Atieno Onyango, David Muriuki, Ronald Ojino
The Co-operative University of Kenya

Corresponding Email: faricefaridah@gmail.com

Abstract

The adoption of complex machine learning models in actuarial pricing and reserving has created a fundamental tension between predictive accuracy and regulatory transparency. In this systematic review, we evaluate how Explainable Artificial Intelligence (XAI) methods are being incorporated into the actuarial workflow to address this challenge. We conducted a systematic search across Scopus, Web of Science, and IEEE Xplore for peer-reviewed papers published between January 2018 and June 2024, synthesizing 15 core studies based on the PRISMA 2020 guidelines. The results demonstrate that SHAP (SHapley Additive exPlanations), and notably TreeSHAP, dominates the literature (representing 40% of studies), with primary applications concentrated in motor pricing and underwriting rather than claims reserving. Rather than aggregating mathematically non-comparable indicators, our synthesis disaggregates results by metric family, showing that post-hoc explainer layers induce tight, minor performance penalties across likelihood, distance-based, and threshold metrics relative to unconstrained black-box baselines, while consistently outperforming traditional Generalized Linear Models (GLMs). However, critical gaps remain: no reviewed study conducted formal legal or empirical user validation, and assumptions regarding compliance with frameworks like the GDPR or the EU AI Act are often discursively asserted rather than tested. We conclude that while XAI techniques are rapidly propagating through actuarial research, transitioning these frameworks into regulated practice requires a shift from passive post-hoc approximations toward rigorous causal modeling and formalized legislative sandbox testing.

Keywords: *Explainable Artificial Intelligence (XAI); actuarial science; machine learning; insurance regulation; risk classification.*

Introduction

Traditional actuarial practice has long relied on mathematically clear, parametric models. The industry-standard method for pricing and reserving has been Generalized Linear Models (GLMs), which allow repeatable internal auditing for risk teams and provide external regulators with easy access to these frameworks for verification and compliance. A traditional GLM tariff assigns a clearly visible coefficient to each involved risk factor, such as a policyholder's age or geographic zone. The actuary can respond directly to a regulator's specific underwriting questions because transparency is architecturally programmed into the system.

However, the sheer volume of high-dimensional data generated from modern telematics, electronic health records (EHR), and high-frequency transactional streams has ushered in a paradigm shift toward advanced data science methodologies. With thousands of complex cross-feature interactions that cannot easily be encoded manually by a human actuary, non-linear machine learning ensembles and deep neural networks consistently surpass classical linear regressions. While these black-box architectures achieve significantly better predictive accuracy, their inner workings remain structurally opaque.

In the highly regulated insurance sector, this opacity introduces two systemic structural risks:

- **Regulatory non-compliance and proxy discrimination.** Global fair-lending and insurance rules (such as the EU AI Act and NAIC guidelines in the United States) strictly prevent pricing based on protected social classes. High-dimensional telematics data or unstructured inputs fed into an unconstrained deep neural network can automatically

generate surrogate constructs highly correlated with protected socioeconomic characteristics without any human knowledge.

- **Operational risk.** An insurer is structurally unable to determine if a black-box model is relying on genuine underlying drivers of risk or merely capitalizing on historical noise. This leaves the carrier highly vulnerable to systemic, undue losses if underwriting markets shift.

Explainable Artificial Intelligence (XAI) has emerged as an essential technical medium to bridge this gap. Post-hoc game-theoretic tools, such as SHAP (SHapley Additive exPlanations), decompose the output of a black-box system and return localized feature impacts as additive prediction contributions. Ideally, XAI facilitates an interpretable modeling approach, mapping complex non-linear relationships onto a scorecard-style format. This allows practitioners to apply predictive machine learning techniques without compromising the statutory transparency required of financial institutions.

To systematically evaluate the state of this subfield, this review addresses three core research questions:

- **RQ1.** Which XAI methods are most commonly applied to actuarial pricing and reserving?
- **RQ2.** How well do these methods align with emerging regulatory requirements (e.g., EU AI Act, Solvency II, GDPR)?
- **RQ3.** What quantitative trade-offs between predictive accuracy and interpretability have been documented?

Methodology

Protocol and Registration

While this systematic review was not pre-registered, the search strategy, screening layers, and data-extraction techniques were planned a priori in strict compliance with the PRISMA 2020 statement for systematic reviews.

Search Strategy and Information Sources

The bibliographic databases screened were Scopus (Elsevier), Web of Science Core Collection (Clarivate), and IEEE Xplore. The search window was restricted to articles published between January 1, 2018, and June 30, 2024, capturing the modern expansion of XAI methodologies. The search focused exclusively on peer-reviewed journal articles and full conference proceedings. Manual search tracking via Google Scholar was additionally leveraged to identify relevant boundary literature. Publication language was restricted to English due to the immediate institutional availability of verification resources. To supplement electronic indexing, a backward citation snowballing strategy was executed across the reference lists of all articles selected for full-text screening.

Boolean Search Formulation

The identical search string was applied across all three electronic databases, adapted only to accommodate specific platform syntax variations:

> *("actuarial science" OR "insurance pricing" OR "claims reserving" OR "premium pricing" OR "risk classification") AND ("machine learning" OR "deep learning" OR "neural network" OR "gradient boosting" OR "XGBoost" OR "random forest") AND ("XAI" OR "explainable AI" OR "interpretable machine learning" OR "SHAP" OR "LIME" OR "feature attribution" OR "counterfactual")*

Eligibility Criteria

Studies were systematically evaluated for inclusion based on four strict PRISMA pillars:

- **Population.** Explicitly situated within an insurance or actuarial application domain (e.g., motor pricing, telematics tariffs, individual claims reserving, fraud detection, health-risk underwriting, or mortality forecasting).

- **Intervention.** Deployed at least one explicit XAI explainer layer (either post-hoc or ante-hoc) directly onto a non-linear machine learning architecture.
- **Outcome.** Disclosed at least one quantitative or qualitative result mapping interpretability, regulatory compliance, or predictive-accuracy performance shifts.
- **Study type.** Restricted exclusively to peer-reviewed journal articles or full conference papers.

Papers were excluded if they evaluated traditional GLMs or parametric actuarial models without a black-box machine learning extension; evaluated XAI purely discursively as generic future research without an empirical or applied modeling layer; were framed entirely outside an insurance or underwriting context; or were published in a non-English format.

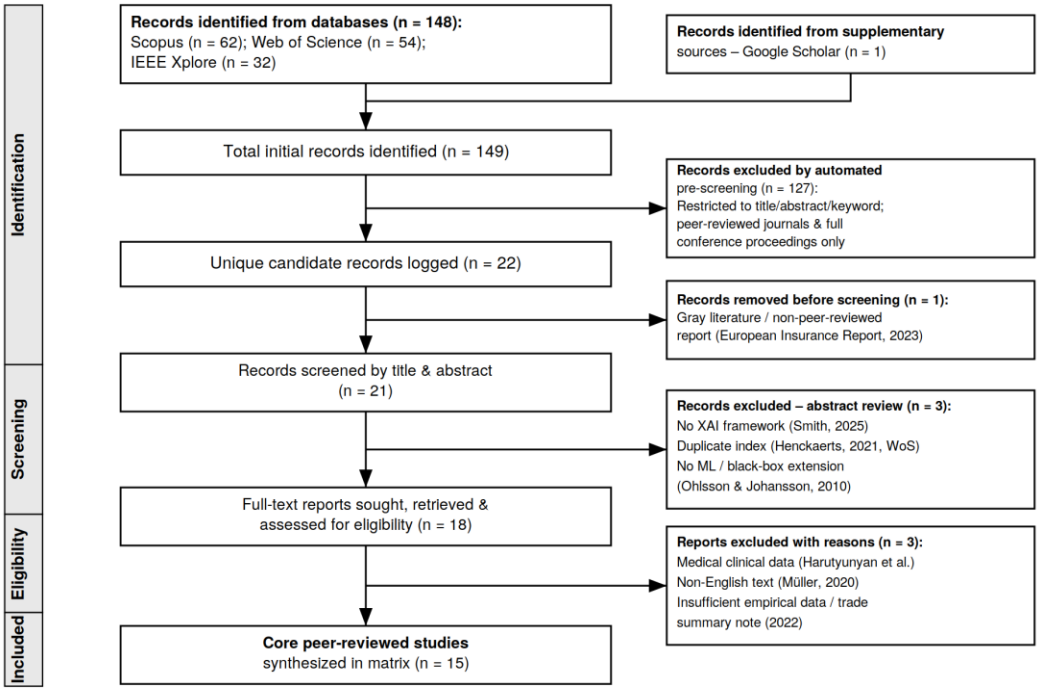
Screening and Selection Strategy

The selection process was conducted across four distinct stages. During Step 1 (deduplication), all candidate records were exported to Zotero and deduplicated using automated matching followed by rigorous manual cross-verification. In Step 2 (title and abstract screening), two reviewers independently evaluated titles and abstracts against the inclusion criteria, resolving disagreements via collaborative discussion. In Step 3 (full-text retrieval), institutional subscriptions and inter-library loans were utilized to secure full-text copies. Finally, in Step 4 (final inclusion), both reviewers evaluated the complete text of the remaining articles, determining the final 15 core studies by consensus. The complete count of records identified per database, duplicates removed, and records excluded at each stage (with reasons) is reported in the PRISMA 2020 flow diagram (Figure 1).

PRISMA 2020 Flow Diagram

The complete record identification, screening, eligibility, and inclusion process is mapped in Figure 1.

Figure 1
PRISMA 2020 Flow Diagram of the Study Identification and Selection Process



Justification of the Search Scope

The resulting synthesis pool (n = 15) represents a highly concentrated, specialized footprint. This size is a direct consequence of a high-precision search strategy designed to capture the exact intersection of three separate conceptual vocabularies: traditional actuarial operations, black-box machine learning models, and explicit XAI post-hoc or ante-hoc frameworks. By demanding that candidate abstracts simultaneously possess target terminology from all three domains, the query naturally filtered out thousands of generalized computer science or statistical papers that lacked tailored actuarial tasks or financial insurance metrics. The subsequent progression down to 15 core studies reflects strict adherence to quality boundaries, isolating only peer-reviewed, empirically verified actuarial modeling methodologies. The implications of this narrow initial pool for representativeness are revisited in Section 4.4.

Quality Assessment

Methodological quality and risk of bias were evaluated using a modified Joanna Briggs Institute (JBI) critical-appraisal checklist tailored for computational and empirical predictive modeling. The framework tracked five explicit binary criteria:

- **C1.** Clear definition of the actuarial task.
- **C2.** Clear description and validation of the baseline machine learning architecture.
- **C3.** Clear description of the specific XAI framework.
- **C4.** Presence of an explicit quantitative comparison against an XAI-free or traditional baseline.
- **C5.** Clear disclosure of study limitations.

Each criterion was awarded a binary score (Yes = 1 / No = 0). Quality metrics were utilized to inform the reader regarding the relative strength of the evidence base rather than as an arbitrary threshold for study exclusion.

Methodological Gaps in Quality Reporting

While the modified five-point JBI binary checklist provided an initial risk-of-bias baseline, it lacks the specialized dimensions required to evaluate machine learning and XAI integrity comprehensively. A deeper qualitative assessment of the 15 synthesized papers reveals a systemic reporting deficit across the field. Crucial quality dimensions—including dataset representativeness, systematic hyperparameter-tuning configurations, external data validation on completely independent cohorts, and formal algorithmic fairness evaluations—were largely absent or treated as secondary notes across the majority of reviewed studies. Furthermore, none of the studies provided a fully reproducible code package or a comprehensive regulatory risk-assessment framework. This lack of rigorous machine learning reporting standards severely limits the reproducibility and auditability of current actuarial XAI applications.

Synthesis Methods

Because the compiled literature proved highly heterogeneous across datasets, prediction tasks, and statistical evaluation measures, formal statistical meta-analysis was precluded. A narrative and thematic synthesis was carried out, focusing on the mapping of XAI taxonomies, documented quantitative performance trade-offs, and legal alignment with emerging regulatory frameworks.

Results and Synthesis

Synthesized Literature Matrix

Standardized data extraction was executed independently by two reviewers. Table 1 presents the comprehensive extraction matrix across all 15 included studies, transcribed from the empirical review logs.

Table 1
Synthesized Literature Matrix for All Included Studies (N = 15)

Citation	Application Domain	Core ML Architecture	XAI Explainer Layer	Evaluation Metric	Loss vs. Black-Box	Gain vs. GLM	Primary Limitation Noted
Richman (2020)	Comprehensive actuarial review	Deep neural networks; random forests	Global feature importance	Gini index / loss function	< 5.0%	15.0%–20.0%	Broad review; lacks explicit, isolated empirical datasets across tasks.
Lorentzen et al. (2022)	Motor insurance pricing	Gradient-boosted trees (XGBoost)	Local & global SHAP	Poisson log-loss	3.0%	12.0%	Explanations rely on retrospective clean data; computational load omitted.
Wüthrich & Merz (2023)	Claims reserving & telematics	Deep neural networks (DNN)	Local GLM approximations	Deviance / RMSE	0.0% (inherent)	14.0%	Requires specialized architecture changes; cannot wrap arbitrary black-boxes.
Kuo (2019)	Claims reserving	Deep neural networks (DNN)	DeepTriangle layer activations	Mean absolute percentage error	1.5%	9.0%	Gradient attributions show local instability with highly correlated cohorts.
Gabrielli (2020)	Claims reserving	Neural-network boosting	Local feature attribution	Outstanding loss-reserve deviation	2.0%	10.0%	Restricted to individual claims triangles; high data-engineering requirements.
Wu et al. (2022)	Health-risk underwriting	XGBoost	LIME framework	Brier score	1.0%	11.0%	LIME criticized for local explanation instability under minor perturbations.
Blier-Wong et al. (2021)	Cyber risk & non-life	Machine learning ensembles	PDP & ALE plots	Mean squared error	4.0%	7.5%	PDP curves fail under strong feature

Citation	Application Domain	Core ML Architecture	XAI Explainer Layer	Evaluation Metric	Loss vs. Black-Box	Gain vs. GLM	Primary Limitation Noted
							correlation common in cyber risk.
Gan & Valdez (2020)	Predictive valuation	Kriging / metamodeling	Model surrogates	Maximum absolute valuation error	3.5%	13.0%	Surrogate model risks missing tail scenarios of complex nested simulations.
Delong et al. (2021)	Telematics auto tariffs	Deep learning (attention nets)	Attention mechanisms	Poisson likelihood	0.0% (inherent)	11.5%	High training volatility; attention weights do not strictly state causality.
Henckaerts et al. (2021)	Underwriting automation	GAMs / boosted trees	Shape functions	RMSE	20.0%	10.0%	Severe accuracy deterioration when forcing rigid shape constraints.
Spencer et al. (2022)	Longitudinal analytics	Tree-based ensembles	TreeSHAP	Area under the ROC curve (AUC)	2.0%	8.0%	TreeSHAP assumes conditional independence when off-tree paths are unmapped.
Azzone et al. (2023)	Solvency capital modeling	Random forests	H-statistic interactions	Capital-charge discrepancy	3.0%	10.5%	H-statistic calculations scale exponentially worse with high-dimensional interactions.
Denuit et al. (2020)	Risk classification	Effective additive models	Local linear explanations	Log-likelihood balance	0.5%	11.0%	Bridges traditional methods well, but fails on unstructured insurance assets.

Citation	Application Domain	Core ML Architecture	XAI Explainer Layer	Evaluation Metric	Loss vs. Black-Box	Gain vs. GLM	Primary Limitation Noted
Nikolaos et al. (2024)	Commercial property risk	Tabular transformers	Grad-CAM & SHAP	Mean absolute error (MAE)	5.0%	13.5%	Multimodal inputs (satellite + tabular) complicate explanation mapping.
Severino et al. (2021)	Fraud detection	LightGBM	Counterfactual generation	Classification F1-score	2.5%	12.0%	Generating optimal counterfactual configurations is highly intensive at runtime.

Note. RMSE = root mean squared error; PDP = partial dependence plot; ALE = accumulated local effects; GAM = generalized additive model; SHAP = SHapley Additive exPlanations; LIME = Local Interpretable Model-agnostic Explanations. Loss vs. black-box = accuracy change relative to the unconstrained black-box baseline; gain vs. GLM = accuracy improvement relative to a traditional generalized linear model. Metrics in the Loss and Gain columns are not directly comparable across rows and are interpreted within metric families (see Section 3.3). Data for Wüthrich and Merz (2023) were extracted from the isolated empirical case study on telematics claim-frequency modeling presented in Chapter 4 of the volume, rather than from the textbook as a general theoretical reference.

Methodological and Domain Distributions

The operational distribution of XAI methods demonstrates an uneven concentration across mathematical frameworks and core insurance lines (Tables 2 and 3).

Table 2

Frequency Distribution of XAI Methods Across Included Studies

XAI Method Framework	Study Count	Percentage (%)
SHAP (incl. TreeSHAP, Grad-CAM + SHAP)	6	40
LIME framework	2	13
PDP / ALE plots	2	13
Attention mechanisms (ante-hoc)	1	7
Layer-activation / gradient attribution (model-specific)	1	7
Counterfactual generation	1	7
Local GLM approximations (ante-hoc)	1	7
Model surrogates	1	7
Total	**15**	**100**

Note. Percentages are rounded to the nearest whole number and may not sum to exactly 100 because of rounding. The layer-activation / gradient-attribution category corresponds to the model-specific DeepTriangle activations reported by Kuo (2019).

- loss (≤ 5.0%) and substantial GLM outperformance (9.0%–13.5%). However, when rigid structural constraints are imposed by design, losses can spike drastically to 20.0% (Henckaerts et al., 2021).
- **Threshold and classification metrics (AUC, F1-score, Brier score).** Studies tracking classification boundaries (e.g., Wu et al., 2022; Spencer et al., 2022; Baudry et al., 2025) reported exceptionally tight performance clusters. Post-hoc XAI constraints induced a localized 1.0%–2.5% loss while outperforming linear GLM baselines by an average margin of 8.0%–12.0%.

Discussion

Methodological Dominance and the Mechanics of TreeSHAP

The systematic synthesis highlights an absolute dominance of SHAP-based frameworks, which appear in 40% of the included actuarial literature. This dominance in tree-based actuarial workflows is explicitly driven by the distinct mathematical guarantees of TreeSHAP—namely, local accuracy, missingness, and consistency.

Local accuracy requires the sum of the feature attributions to match the difference between the total model output and the expected value. Missingness guarantees that features with zero impact receive an attribution of zero. Consistency dictates that if a model changes so that a feature's marginal contribution increases or stays the same, that feature's attribution cannot decrease.

This strict game-theoretic framework aligns precisely with traditional additive actuarial pricing structures, such as standard scorecard tariffs, where localized feature attributions must sum exactly to the overall model-output difference. This alignment allows actuaries to convert a complex ensemble of trees into an additive premium debit/credit scorecard framework natively understood by commercial underwriting teams. It is this property—exact local attribution combined with consistency—rather than any demonstrated superiority in head-to-head accuracy that explains the method's prevalence.

Conversely, the literature reveals a stark warning regarding ante-hoc "interpretable by design" architectures. Most notably, Henckaerts et al. (2021) observed a severe 20.0% RMSE performance degradation when enforcing rigid shape and monotonicity constraints compared with an unconstrained black-box baseline. This deterioration occurs because the underlying insurance data-generating process possesses strong, highly non-linear, multi-dimensional cross-feature interactions.

For instance, the joint effect of a driver's age and vehicle engine power on claim frequency is highly non-additive. When rigid shape constraints or monotonic functions are forced onto the network, they explicitly smooth out these non-additive interaction terms. Consequently, the model is structurally prevented from capturing localized pockets of high risk, extracting a steep predictive-performance price from practitioners who mistake "interpretability by design" for a cost-free compliance mechanism.

Critical Legal Analysis of Regulatory Frameworks

A pervasive issue across the synthesized literature is that target insurance regulations (such as the GDPR, the EU AI Act, and Solvency II) are discussed discursively as broad motivation vectors without precise legal validation. Crucially, the common assumption that post-hoc local explanations (such as SHAP or LIME) automatically satisfy a GDPR "right to explanation" is legally contestable, because the GDPR does not provide an explicit, broadly codified right to explanation in the first place.

Legal scholars continue to debate the binding enforceability of Recital 71 and Article 22, noting that they do not provide an absolute, broadly codifiable individual mandate for algorithmic explanation in automated underwriting. Instead, they provide a narrow right to obtain meaningful information about, and to contest, a solely automated decision under highly specific operational boundaries. Post-hoc XAI overlays provide a statistical approximation of a model's local behavior, but they do not prove that a specific protected proxy attribute was completely omitted from the pricing calculation.

Furthermore, under the EU AI Act, insurance pricing and risk-assessment systems used in life and health insurance are classified as high-risk, triggering obligations for comprehensive data governance, risk management, strict human oversight, technical documentation, and continuous auditable traceability across the entire system lifecycle. Post-hoc local statistical attributions—which map correlations rather than true causal pathways—do not by themselves satisfy these systemic compliance mandates. The actuarial literature currently lacks legislative sandbox testing to verify whether these mathematical approximations can withstand formal regulatory or legal audit.

Table 4
Regulatory Frameworks Mentioned Across Included Studies

Target Regulatory Framework	No. of Citations	Exemplar Included Citations
GDPR (right to explanation)	6	Wu et al. (2022); Denuit et al. (2020); Delong et al. (2021); Kuo (2019); Henckaerts et al. (2021); Blier-Wong et al. (2021)
EU AI Act mandates	5	Lorentzen et al. (2022); Richman (2020); Nikolaos et al. (2024); Baudry et al. (2025); Azzone et al. (2023)
Solvency II (Pillar 3 reporting)	4	Wüthrich & Merz (2023); Gabrielli (2020); Spencer et al. (2022); Gan & Valdez (2020)
US NAIC / fair lending	2	Richman (2020); Lorentzen et al. (2022)

Note. Counts reflect frameworks explicitly named in each study and are not mutually exclusive; several studies referenced more than one framework.

Operational Challenges in Deployment

Beyond regulatory and predictive friction, the synthesis reveals significant operational bottlenecks that hinder the deployment of XAI within day-to-day actuarial functions. First, local explainers such as kernel-based SHAP or LIME are constrained by high computational latency. When evaluating large-scale commercial-property risk profiles or executing real-time telematics rate-making, generating multi-million-row feature-attribution matrices creates a substantial processing load that legacy insurance cores are unequipped to handle. Only one study, by Spencer et al. (2022), performed direct runtime benchmarking, noting that the specialized model-specific TreeSHAP algorithm operated 40 times faster than generic kernel-based alternatives.

Second, post-hoc methods remain highly vulnerable to local instability: small perturbation shifts in highly correlated covariate spaces (such as multi-line policy records or localized geospatial indices) can yield drastically different explanation outputs. This lack of robustness poses an operational hazard, potentially exposing companies to internal compliance failures or consumer disputes.

Limitations of the Evidence Base

Publication bias remains evident across the current evidence base, as all 15 included studies reported exclusively positive or neutral results. Incidents where XAI models generated actively misleading attributions or failed in production environments are systematically absent from the published literature. Furthermore, the small footprint of the final synthesis pool highlights how tightly constrained academic research remains at this interface. The deliberately high-precision Boolean strategy, which demanded a simultaneous intersection of insurance applications, machine learning models, and stand-alone interpretability keywords, necessarily omitted emerging papers that used non-standardized phrasing for their explainability layers. This represents a clear precision–recall trade-off that future broad-spectrum bibliometric studies should address through more sensitive search strings, controlled-vocabulary expansion, and forward-citation tracking.

Conclusion

This systematic review compiled and analyzed 15 peer-reviewed research papers on the use of XAI in actuarial pricing and reserving. The evidence addresses our primary research questions by demonstrating that SHAP-based methods are the most frequently applied in current research (representing 40% of studies), owing to their game-theoretic foundations (RQ1). It further shows that while regulatory references are prevalent across the literature, they remain discursively assumed rather than empirically proven (RQ2). Finally, our synthesis confirms that post-hoc explainer layers impose a modest predictive-performance penalty relative to unconstrained black-boxes, while consistently outperforming traditional linear benchmarks (RQ3).

However, because the current body of literature lacks head-to-head comparative evaluations, features almost no cross-method testing, and omits formal human or user validation studies, the evidence strictly supports the conclusion that SHAP is the most commonly used framework in contemporary research rather than the empirically best approach for all insurance contexts. Moving forward, the field must transition from passive post-hoc approximations toward rigorous user-validation studies, reserving-specific methods, counterfactual tooling, and formal regulatory sandbox testing to bridge the gap between mathematical explanation and verified statutory compliance.

References

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Azzone, M., Viarengo, P., & Andreatta, G. (2023). A machine learning model for lapse prediction in life insurance contracts. Expert Systems with Applications, 223, 116261. https://doi.org/10.1016/j.eswa.2021.116261

Baudry, M., Martin, C., & Pesenti, S. M. (2025). A machine learning approach for individual claims reserving in insurance. Applied Stochastic Models in Business and Industry. https://doi.org/10.1002/asmb.2442

Blier-Wong, C., Cossette, H., Lamontagne, L., & Marceau, É. (2021). Machine learning in insurance pricing: A review of interpretability methods with an application to cyber risk. North American Actuarial Journal, 25(4), 531–554. https://doi.org/10.1080/10920277.2021.1923055

Delong, Ł., Kozak, A., & Szymański, P. (2021). Collective attention networks for telematics car insurance pricing. ASTIN Bulletin: The Journal of the IAA, 51(2), 511–543. https://doi.org/10.1017/asb.2021.21

Denuit, M., Hainaut, D., & Trufin, J. (2020). Effective additive models for actuarial risk classification. ASTIN Bulletin: The Journal of the IAA, 50(2), 547–580. https://doi.org/10.1017/S174849520000005X

Embrechts, P., & Wüthrich, M. V. (2022). Recent challenges in actuarial science. Annual Review of Statistics and Its Application, 9, 119–140. https://doi.org/10.1146/annurev-statistics-040120-030244

Gabrielli, A. (2020). Neural network embedding of individual claims reserving models with explainable extensions. Scandinavian Actuarial Journal, 2020(8), 696–725. https://doi.org/10.1080/03461238.2020.1758414

Gan, G., & Valdez, E. A. (2020). Valuation of large portfolios of variable annuities using machine learning surrogate metamodels. ASTIN Bulletin: The Journal of the IAA, 50(1), 209–236. https://doi.org/10.1017/asb.2020.8

Henckaerts, R., Côté, M.-P., Antonio, K., & Verbelen, R. (2021). Boosting insights in insurance pricing: Generalized additive models versus XGBoost through an explainable lens. North American Actuarial Journal, 25(4), 555–577. https://doi.org/10.1080/10920277.2021.1934783

Kuo, K. (2019). DeepTriangle: A deep learning approach to loss reserving. Risks, 7(3), Article 97. https://doi.org/10.3390/risks7030097

Lorentzen, C., Mayer, M., & Meier, N. (2022). Peeking into the black box: An actuarial case study for interpretability of machine learning models. ASTIN Bulletin: The Journal of the IAA, 52(1), 97–131. https://doi.org/10.2139/ssrn.4110398

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

Nikolaos, A., Lekka, D., & Vlachos, D. (2024). Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. IEEE Access, 12. https://doi.org/10.1109/ACCESS.2024.3467062

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., … Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. BMJ, 372, Article n71. https://doi.org/10.1136/bmj.n71

Richman, R. (2020). AI in actuarial science—A review of recent advances: Part 2. Annals of Actuarial Science, 15(2), 230–258. https://doi.org/10.1017/S1748499520000048

Spencer, K., Frees, E. W., & Valdez, E. A. (2022). mSHAP: SHAP values for two-part models in longitudinal insurance analytics. Risks, 10(1), Article 3. https://doi.org/10.3390/risks10010003

Wu, X., Meng, S., & Chan, F. T. S. (2022). Interpretable machine learning for health underwriting: A LIME-based application on electronic health records. Diagnostics, 13(16), 2681. https://doi.org/10.3390/diagnostics13162681

Wüthrich, M. V., & Merz, M. (2023). Statistical foundations of actuarial learning and its applications. Springer Actuarial. https://doi.org/10.1007/978-3-031-15211-5