A Segmentation-Based Hybrid Chain Ladder–XGBoost Framework for Incurred But Not Reported Claims Estimation

N. M. Ajang[1*], O. Ngesa[2], J. Aduda[3]

[1*]Pan African University Institute for Basic Sciences, Technology and Innovation (PAUSTI), Kenya
[2]Taita Taveta University, Kenya.
[3]Jomo Kenyatta University of Agriculture and Technology (JKUAT), Kenya.

ABSTRACT: Accurate estimation of Incurred But Not Reported (IBNR) claims is critical in non-life insurance, particularly for portfolios characterized by heterogeneous claim sizes and nonlinear development patterns. Existing hybrid reserving models improve predictive accuracy but typically apply multiple models uniformly to the same claim population, implicitly assuming homogeneous claim behavior. Under heterogeneous portfolios, large and volatile claims can simultaneously distort Chain Ladder development factors and destabilize machine-learning training. This study addresses this limitation by proposing a segmentation-based hybrid Chain Ladder (CL)–XGBoost framework that models claim according to their volatility characteristics rather than treating all claims identically. Claims are partitioned into small and large segments using a data-driven 65th percentile threshold (claim amount €32.49). The threshold is selected through sensitivity analysis by minimizing validation-set Mean Absolute Error (MAE) prior to test evaluation, ensuring that it is not arbitrary or tuned to the test data. Small claims, which exhibit stable development patterns, are modeled using the classical CL method, whereas large high-variance claims are modeled using XGBoost, and segment-level predictions are aggregated to obtain cumulative IBNR estimates. Using motor insurance claims data from 2015–2019, the proposed framework demonstrates substantial improvements over benchmark models. Compared with the classical CL baseline (MAE €1,774.51), the hybrid model reduces MAE to €819.28 (54% reduction), RMSE from €2,026.08 to €819.39 (60% reduction), and MAPE from 0.92% to 0.60%, relative to an average claim size of approximately €387. Bootstrap resampling confirms that the reductions are statistically significant at the 5% level. The proposed model also outperforms standalone XGBoost, Uniform Hybrid and Neural-Network–CL hybrid across all evaluation metrics. The results show that performance gains arise primarily from heterogeneity-aware segmentation rather than hybridization alone. The framework therefore provides a robust and interpretable reserving approach that stabilizes development estimation for small claims while allowing flexible nonlinear learning for large claims.

KEYWORDS: Chain Ladder, Hybrid Actuarial Model, IBNR, Insurance Reserving, Machine Learning, XGBoost.

[Received: Feb. 11, 2026; Revised: Feb. 22, 2026; Accepted: Mar. 17, 2026] Print ISSN: 0189-9546 | Online ISSN: 2437-2110

I. INTRODUCTION

Over the last decade, machine learning (ML) techniques have undergone an extraordinary surge in popularity and application across many fields of science and industry. The growing ability to process large volumes of data, together with advances in algorithmic methods, has accelerated their adoption. The insurance industry, however, has historically adopted ML more slowly due to regulatory constraints, limited data availability, privacy concerns, and limited familiarity among actuaries with advanced analytics techniques (Hiabu et al., 2023). As regulatory barriers diminish and data availability improves, interest in ML applications for underwriting and claims reserving has increased substantially (Mahohoho et al., 2023; Blier-Wong et al., 2020; Kuo, 2019; Avanzi et al., 2024). One major advantage of ML techniques over traditional actuarial models, particularly generalized linear models (GLMs), is their ability to capture complex nonlinear relationships and interaction effects without requiring the analyst to pre-specify functional forms. This flexibility allows ML models, such as gradient boosting methods, to adapt to heterogeneity in claim frequency, severity, and reporting patterns. Nevertheless, ML performance remains sensitive to data quality, model tuning, and structural changes in the data-generating process (Manathunga and Zhu, 2022; Song and Heo, 2022; Poufinas et al., 2023; Chen, 2016). The increasing use of alternative data sources, including telematics and unstructured textual information, further strengthens the relevance of ML-based reserving approaches (Hiabu et al., 2023).

Despite the advantages of ML methods, traditional reserving techniques remain dominant in actuarial practice. The CL method is the most widely used approach for estimating non-life insurance claim reserves, particularly incurred but not reported (IBNR) liabilities, due to its conceptual simplicity, transparency, and regulatory acceptance. The method uses cumulative development patterns summarized in run-off triangles to estimate ultimate losses. However, although non-parametric, the CL approach relies on strong assumptions regarding stability of claim development and independence of development periods. In heterogeneous

*Corresponding author: *ajang.nhial@students.jkuat.ac.ke*

doi: http://doi.org/10.63746/njtd.v23i1.4437

portfolios, these assumptions may be violated, leading to unstable development factors and distorted reserve projections, especially when a small number of large claims dominates development patterns (Saoudi et al., 2023; Suwardi and Purwono, 2021; Krisdiantha et al., 2023; Mazzi et al., 2024). Accurate reserve estimation is fundamental to insurer solvency and financial stability because reserves determine the funds set aside to meet future claim obligations and directly influence financial reporting and capital adequacy. IBNR reserves complement reported but not settled (RBNS) reserves and are particularly sensitive to modeling assumptions, reporting delays, and claim heterogeneity. Consequently, recent research emphasizes micro-level modeling and ML-based reserving approaches to improve predictive accuracy (Hiabu and Müller, 2021; Bücher and Rosenstock, 2023; Avanzi et al., 2023; Henckaerts et al., 2021; Antonio and Plat, 2021).

Although ML models such as XGBoost and Random Forests often outperform traditional actuarial methods in complex reserving environments, empirical findings remain mixed. In certain datasets and portfolio segments, the CL method performs comparably to or better than ML models (Blier-Wong et al., 2020). This reflects a practical bias–variance trade-off: the CL method may exhibit low bias under stable development but high variance under heterogeneity and outliers, whereas flexible ML models reduce bias by capturing nonlinear patterns but may display higher variance and lower interpretability if unconstrained (Dong and Quan, 2025; Laub et al., 2025). To address these limitations, hybrid reserving frameworks have been proposed. Wüthrich (2018) introduced a neural-network extension of the CL method in which development factors depend on claim-specific characteristics. Other studies proposed robust CL variants using median-based development factors or adjustments for extreme observations (Feng and Li, 2023; Suwardi and Purwono, 2021), while segmentation approaches based on clustering revealed heterogeneous development patterns and motivated segment-specific reserving strategies (Saida et al., 2023). Segmentation has been increasingly recognized as an effective strategy for handling portfolio heterogeneity in claims reserving, particularly when claim size distributions are heavy-tailed and large claims disproportionately influence development factor estimation. By separating claims with stable development behavior from those exhibiting high volatility, segmentation can improve the stability of traditional reserving methods while allowing more flexible models to capture complex patterns in volatile segments**.** However, most hybrid frameworks still apply models uniformly across the entire portfolio and therefore do not fully exploit differences between stable and volatile claims.

Insurance portfolios typically contain predictable claims that follow stable development patterns alongside high-variance large claims that distort development factor estimation and ML training. This suggests that reserving performance may improve when modeling strategies are adapted to claim characteristics rather than applied uniformly. Accordingly, this study proposes a segmentation-based hybrid reserving framework that applies the CL method to stable claims and the XGBoost algorithm to volatile claims. Claims are partitioned using a percentile-based segmentation threshold, where claims above the 65th percentile (€32.49) are classified as large and modeled with XGBoost, while smaller claims are modeled using CL. The segmentation rule is empirically evaluated by comparing predictive performance against non-segmented benchmark approaches, including standalone XGBoost, classical CL, and a Neural Network–Chain Ladder hybrid model. Claim heterogeneity directly affects the stability of CL development factors because large claims disproportionately influence cumulative development ratios, while ML models applied uniformly may overfit numerous small claims and exhibit high prediction variance. The proposed segmentation framework addresses this bias–variance trade-off by stabilizing development factor estimation within homogeneous small-claim groups while allowing flexible nonlinear learning for volatile large claims. To isolate the value of segmentation itself, the segmented hybrid framework is also evaluated against a non-segmented hybrid model in which both CL and XGBoost are applied to the full dataset without claim partitioning.

The contributions of this study are threefold. First, it introduces a percentile-based segmentation rule that isolates high-variance large claims without removing valid observations. Second, it proposes a segment-specific hybridization strategy combining CL and XGBoost and aggregates segment-level predictions into cumulative IBNR estimates. Third, empirical analysis using motor insurance data from 2015–2019 demonstrates statistically and economically meaningful improvements in predictive accuracy relative to benchmark methods. The remainder of the paper is organized as follows: Section II presents the theoretical and conceptual framework, Section III describes the data and methodology, Section IV presents the results and discussion, and Section V concludes the paper with implications and directions for future research.

II. THEORETICAL AND CONCEPTUAL FRAMEWORK

The theoretical foundation of insurance claims reserving is grounded in actuarial science, where the objective is to estimate future liabilities arising from past insured events. In non-life insurance, a substantial portion of total reserves consists of Incurred But Not Reported (IBNR) claims, representing losses that have occurred but have not yet been reported. Accurate estimation of IBNR reserves is essential for solvency assessment, capital adequacy, and financial stability. Traditional actuarial reserving techniques rely on structured statistical assumptions and aggregated loss development data to project future claims development. Among these methods, the CL technique remains the most widely used methodology because of its conceptual simplicity, transparency, and regulatory acceptance (SAOUDI et al., 2023; Suwardi and Purwono, 2021; Krisdiantha et al., 2023; Mazzi et al., 2024).

The CL method operates on cumulative claims development triangles and assumes that historical development patterns are predictive of future development. Development factors estimated from observed data are applied to incomplete accident years to forecast ultimate claims. Although the CL method is non-parametric, it relies on strong assumptions, including the stability of development factors across time and

the independence of development periods. These assumptions may be violated in heterogeneous or rapidly changing portfolios, particularly when a small number of large claims dominate development patterns and distort estimated development factors (SAOUDI et al., 2023; Suwardi and Purwono, 2021). Consequently, purely aggregate reserving approaches may produce unstable reserve estimates in the presence of claim heterogeneity. Recent research has therefore emphasized the limitations of purely aggregate reserving techniques and highlighted the need for more flexible modeling approaches in complex reserving environments (Hiabu and Müller, 2021; Bücher and Rosenstock, 2023; Henckaerts et al., 2021; Antonio and Plat, 2021).

In contrast to traditional actuarial techniques, ML methods provide a flexible modeling framework capable of capturing nonlinear relationships and interaction effects without requiring strong parametric assumptions. While generalized linear models (GLMs) require the analyst to specify the functional relationship between predictors and the response variable, ML algorithms learn patterns directly from data. Tree-based ensemble methods such as Extreme Gradient Boosting (XGBoost) are particularly suitable for insurance applications because they can model nonlinear claim development dynamics and automatically account for complex interactions among variables (Chen, 2016; Rahul, 2020; Mahohoho et al., 2023). Empirical studies have shown that ML models can outperform traditional actuarial techniques in long-tail or nonlinear reserving contexts (Rahul, 2020; Mahohoho et al., 2023; Avanzi et al., 2023).

However, the superiority of ML models is not universal. Their performance depends heavily on data quality, feature engineering, and hyperparameter tuning, and they may not generalize well under structural changes (Manathunga and Zhu, 2022; Song and Heo, 2022; Poufinas et al., 2023). Moreover, concerns regarding interpretability and alignment with actuarial standards have been raised, potentially limiting adoption in regulated insurance markets (Dong and Quan, 2025; Laub et al., 2025). Some empirical studies report that traditional actuarial methods, particularly the CL approach, can perform comparably or even better than ML models in certain datasets or claim segments (Blier-Wong et al., 2020). This reflects a practical bias–variance trade-off: traditional methods may be stable under homogeneous development but sensitive to outliers, whereas flexible ML models can capture nonlinearities but may exhibit higher variance when applied to stable claim patterns.

Recognizing the complementary strengths and weaknesses of traditional actuarial methods and ML approaches, recent research has increasingly focused on hybrid reserving frameworks. One notable contribution is the neural network–based extension of the CL method, which models development factors as functions of claim-specific characteristics while preserving actuarial structure (Wüthrich, 2018). Other studies have developed robust versions of the CL method that reduce sensitivity to outliers through median-based development factors or residual-based adjustments (Feng and Li, 2023; Suwardi and Purwono, 2021). Empirical evidence suggests that these hybrid and robust approaches can improve predictive performance in datasets containing extreme claims, demonstrating the value of combining actuarial structure with learning-based or robustness components.

Another important conceptual element in modern reserving theory is claim heterogeneity. Insurance portfolios often contain claims with widely varying sizes, development speeds, and volatility characteristics. Aggregate modeling approaches that apply a single method uniformly across all claims may fail to capture this heterogeneity because large claims can disproportionately influence development factors and reserve projections. Segmentation techniques, including clustering of run-off triangle data, have therefore been proposed to identify groups of claims with similar development behavior and to inform tailored modeling strategies (Saida et al., 2023). This line of research supports the principle that no single model is universally optimal for all claim types, particularly in heterogeneous or long-tailed datasets (Blier-Wong et al., 2020).

Building on these theoretical perspectives, the conceptual framework of this study is based on segmentation-driven hybrid modeling. The central premise is that stable, low-variance claims tend to follow predictable development patterns that are well captured by the CL method, while large or high-variance claims exhibit nonlinear and irregular dynamics that are more appropriately modeled using flexible ML algorithms such as XGBoost. Claims are partitioned according to claim amount using a percentile-based rule. Specifically, a claim is classified as a large claim if its value exceeds the 65th percentile of the claim size distribution (threshold €32.49), while claims below this threshold are classified as small claims. This threshold was selected through validation-based sensitivity analysis to reduce within-segment heterogeneity without removing valid observations.

Within this structure, the CL method is applied to the stable segment to leverage its interpretability and regulatory familiarity, while XGBoost is employed for the volatile segment to capture nonlinear relationships and interaction effects (Chen, 2016; Wüthrich, 2018). Segment-level predictions are subsequently aggregated to obtain cumulative IBNR reserve estimates. This integration reflects a synthesis of actuarial loss development theory and statistical learning theory, aiming to improve predictive accuracy while maintaining transparency and practical relevance in insurance reserving.

III. MATERIALS AND METHOD

A. The Study Dataset.

This study utilizes an anonymized motor vehicle insurance claims dataset covering the period 2015–2019. The dataset contains 109,188 individual claim records, each representing a distinct insurance claim processed through the insurer's digital claims management platform. All personally identifiable information was removed prior to analysis, and the dataset was used solely for academic research purposes. The records include claim payment information, policy characteristics, and vehicle attributes. The portfolio consists of both private and commercial vehicles, introducing variation in claim frequency, claim severity, and reporting behavior. Such variability is particularly relevant for Incurred But Not

Reported (IBNR) estimation, as heterogeneous development patterns may influence reserve adequacy.

The temporal coverage of the dataset allows the construction of cumulative claim development triangles and analysis of reporting patterns. Consequently, the dataset provides a suitable empirical basis for comparative evaluation of traditional actuarial reserving techniques, such as the CL method, and data-driven approaches, including Extreme Gradient Boosting (XGBoost).

B. Data Used and Sources.

The dataset contains both claim-level and policy-level variables describing claim development, policy exposure, and insured vehicle characteristics. Table 1 presents a summary of the variables and their descriptions

Table 1 Summary of the key variables and descriptions.

No.	Variable	Description
1	ID	Unique identifier for each policyholder
2	PolicyYearKey	Policy year key
3	Claim_ID	Unique identifier for each claim
4	Accident_Date	Date of accident occurrence
5	Accident_Year	Year of accident occurrence
6	Claim_Total	Total claimed amount for the accident
7	Incremental_Paid	Incremental claim payment in the period
8	Cumulative_Paid	Cumulative claim payments to date
9	Exposure_Days	Number of active policy days in the period
10	Date_last_renewal	Last policy renewal date
11	Date_next_renewal	Next scheduled renewal date
12	Distribution_channel	Policy acquisition mode (online, agent, broker)
13	Seniority	Years insured with the company
14	Policies_in_force	Number of active policies
15	Max_policies	Maximum allowed policies
16	Max_products	Maximum product coverage allowed
17	Lapse	Indicator of policy lapse/cancellation
18	Payment	Claim payment mode or status
19	Premium	Policy premium amount
20	Cost_claims_year	Total annual claim cost
21	N_claims_year	Number of claims in current year
22	N_claims_history	Historical number of claims
23	R_Claims_history	Historical claim ratio (risk indicator)
24	Type_risk	Type of insured risk
25	Area	Geographical region
26	Second_driver	Indicator for second driver coverage
27	Year_matriculation	Vehicle registration year
28	Power	Engine power
29	Cylinder_capacity	Engine capacity (cc)
30	Value_vehicle	Vehicle market value
31	N_doors	Number of vehicle doors
32	Type_fuel	Type of fuel used
33	Length	Vehicle length (meters)
34	Weight	Vehicle weight (kilograms)

For the ML model, the feature vector x consists of both claim-level and policy-level predictors derived from Table 1. These include claim history indicators (number of prior claims and claim ratios), exposure variables (exposure days and premium), policy attributes (renewal dates and coverage characteristics), and vehicle characteristics (vehicle value, power, age, fuel type, and engine capacity). This specification allows XGBoost to model nonlinear relationships between observable risk characteristics and claim development while maintaining consistency with actuarial reserving structures.

C. Data collection and processing

The dataset was obtained in CSV format from the insurer's online claims platform. Each record corresponds to an individual claim and includes details such as Claim_Total, Incremental_Paid, Cumulative_Paid, and policy-related information. However, certain temporal variables necessary for claims development studies, specifically the accident date and reporting date, were not available in the original dataset. These variables are required to construct incremental and cumulative run-off triangles and to estimate Incurred But Not Reported (IBNR) reserves. Consequently, accident and reporting dates were generated synthetically, resulting in a database that combines observed claim severities with simulated claim timing information.

The simulation was limited strictly to timing variables only. All claim payment amounts and policy characteristics remain the original observed insurance data. Therefore, the reserving structure of the dataset is not imposed by any actuarial or machine learning reserving method, and the simulation does not structurally favor either the CL or XGBoost model. Accident dates were generated using a Poisson arrival process to reproduce realistic claim occurrence patterns while matching the observed number of claims per accident year. Only the training accident years (2015–2016) were used to calibrate the arrival intensity in order to prevent information leakage from validation and test periods. Reporting delays were not observable in the dataset and therefore could not be statistically estimated. Instead, reporting delays were modeled as lognormal random variables calibrated to actuarial reporting patterns commonly observed in motor insurance. A lognormal distribution with parameters $\mu = 1.2$ and $\sigma = 0.7$ was selected to produce a realistic mean reporting delay of approximately 3–6 months, consistent with reporting delay patterns discussed in stochastic reserving literature (England and Verrall, 2002; Wüthrich and Merz, 2008). The parameters were calibrated to represent plausible industry reporting behavior rather than fitted to the dataset. Claim severity and reporting delay were assumed independent, which is a common simplifying assumption in claims reserving models when delay information is unavailable (Antonio and Plat, 2021). Severity values were preserved from observed data and not generated from the delay distribution.

To evaluate robustness, sensitivity analysis was conducted using alternative variance values for the lognormal delay distribution. Model performance was compared across low-, medium-, and high-variance delay scenarios, and the relative performance ranking of the reserving models remained stable.

Claim severity was not simulated. Instead, the observed claim amounts were preserved and linked to simulated reporting delays. For development consistency, severity allocation across reporting periods was modeled using a gamma distribution calibrated from the training-set payment patterns. This allows heterogeneity in claim size to remain present in the data while introducing realistic development timing. Because the claim amounts themselves are not generated by the reserving models, the simulation does not embed development factors or favor either traditional actuarial or machine learning approaches. The purpose of the synthetic timing variables is therefore not to reproduce historical claim histories exactly, but to create internally consistent development prcesses that allow fair comparison of reserving methodologies under realistic reporting conditions.

Outliers in the target variable Incremental_Paid were retained. Extreme values represent genuine high-severity claims characteristic of insurance portfolios and typically follow heavy-tailed distributions. Removing such claims would bias reserve estimation and artificially stabilize development factors. Retaining them ensures that reserve estimates remain representative of real insurance risk.

D. Test design

The reserving analysis adopts a valuation-based framework consistent with actuarial practice. All claims are observed up to a fixed valuation age of 12 months of development, which represents the information available to an insurer at the valuation date. The ultimate claim amount is approximated at 36 months of development. Therefore, the portion of development occurring between 12 and 36 months is treated as the unobserved Incurred But Not Reported (IBNR) component.

The selection of the 36-month horizon is based on empirical development completeness of the portfolio. An analysis of cumulative paid losses for the training accident years was performed to evaluate payment maturity by development age. The cumulative paid proportion increases rapidly during the early development periods and stabilizes near 36 months, with only minimal incremental payments observed thereafter. Consequently, the 36-month development age is considered a practical approximation of the ultimate claim value for this motor insurance portfolio, and the development between 12 and 36 months represents the reserve to be predicted. In actuarial reserving practice, ultimate development is commonly approximated at the point where incremental payments become negligible relative to cumulative losses. The observed stabilization of cumulative payments around 36 months indicates that the majority of claim settlement activity has occurred, making it a practical proxy for ultimate loss in this portfolio.

To evaluate predictive performance under realistic reserving conditions, the dataset is partitioned chronologically by accident year. Accident years 2015–2016 are used for model training, accident year 2017 is used for model validation and model selection, and accident years 2018–2019 are reserved for out-of-sample testing. This chronological separation ensures that future claim development information is not used in model estimation and prevents information leakage across development periods.

All reserving methods are calibrated using only information available at the 12-month valuation point. Development information beyond 12 months is not used during model training or model selection. For the machine learning models, incremental payments beyond the valuation age are predicted at the individual-claim level and subsequently aggregated by accident year and development period to form projected run-off triangles. This aggregation ensures methodological comparability between traditional actuarial reserving approaches and micro-level machine learning

predictions while maintaining consistency in reserve estimation.

This experimental design replicates the operational reserving environment in which insurers must forecast future claim development using incomplete information at a fixed valuation date. Although the number of test accident years is limited, evaluation at the accident-year level is standard in actuarial reserving because reserve adequacy is assessed at the portfolio level rather than through large-sample statistical inference. Each accident year represents a complete reserving problem, and predictive performance is evaluated on aggregate reserve quantities. Therefore, out-of-sample validation across multiple accident years provides an operationally relevant assessment of model performance.

E. Claim Amount—Based Segmentation

To address heterogeneity in claim development behavior and to reduce distortion of development factors by extreme severities, the dataset is segmented based on cumulative claim amount observed at the valuation date. Smaller claims generally exhibit stable development patterns, while large claims display greater volatility and nonlinear behavior.

Let y_i denote the cumulative claim amount for claim i at the valuation date and let τ denote the segmentation threshold. Claims are classified as follows:

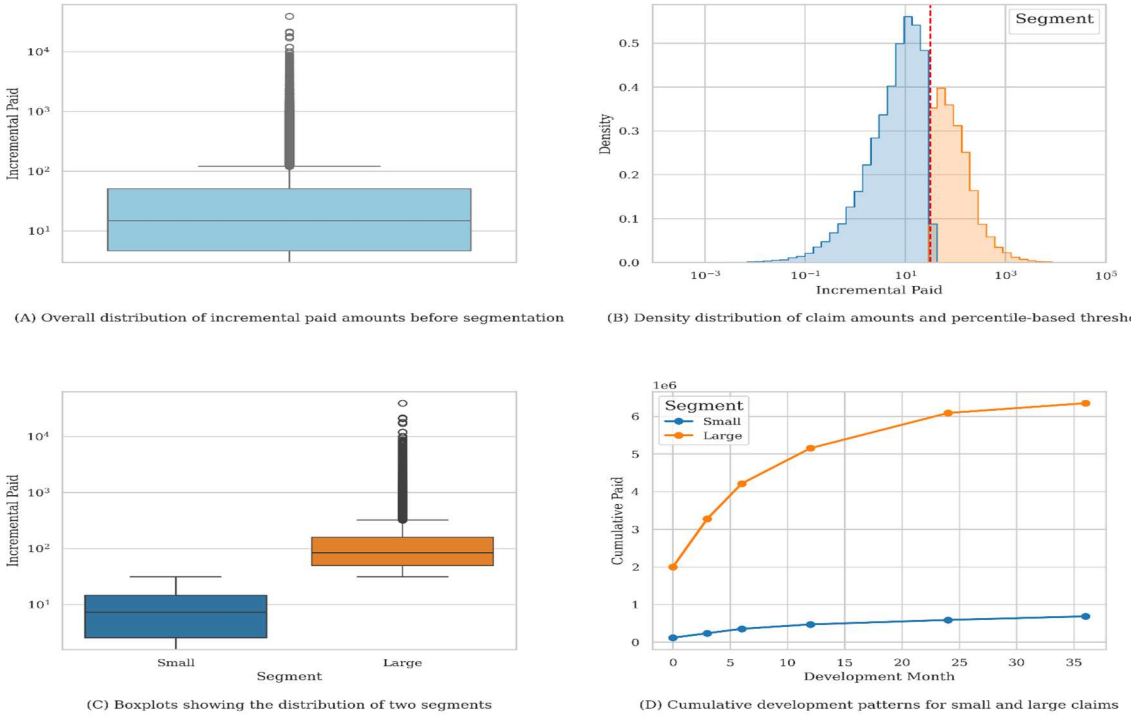
Small claims: $y_i \leq \tau$ (1)

Large claims: $y_i > \tau$ (2)

The threshold τ is selected using a percentile-based, data-driven procedure. Candidate thresholds corresponding to the 60th, 63rd, 65th, 67th, and 70th percentiles were evaluated on the validation dataset. The optimal threshold was defined as the value that minimized the validation Mean Absolute Error (MAE) of the hybrid model prior to test evaluation. The 65th percentile (τ = €32.49) produced the lowest validation error and was therefore selected. Importantly, the threshold selection was completed using only the validation year (2017). The test years (2018–2019) were not used during threshold determination, ensuring that the segmentation rule was not tuned to out-of-sample results.

Small claims are modeled using the CL method to leverage stable development behavior, while large claims are modeled using XGBoost to capture nonlinear dynamics and interaction effects. Segment-level predictions are then aggregated to obtain total IBNR reserve estimates. This segmentation stabilizes development factor estimation for the actuarial component while allowing ML to model high-variance claim behavior.

Figure 1 illustrates the claim amount–based segmentation process. Panel (A) shows the distribution of cumulative claim amounts at the valuation date prior to segmentation, Panel (B) presents the selected percentile threshold, Panel (C) shows the resulting segment distributions, and Panel (D) displays the cumulative development patterns for each segment.



(A) Overall distribution of incremental paid amounts before segmentation

(B) Density distribution of claim amounts and percentile-based threshold

(C) Boxplots showing the distribution of two segments

(D) Cumulative development patterns for small and large claims

Figure 1 Empirical distribution of cumulative claim amounts before and after segmentation

F. Model Development Framework

This section presents the modelling framework adopted in this study. We first introduce the classical (CL) method, followed by the XGBoost model. Finally, we present the proposed Hybrid CL–XGBoost framework, which integrates both approaches for improved IBNR estimation.

G. The CL Method

The CL method is a classical actuarial reserving technique used to estimate ultimate claims and incurred but not reported (IBNR) liabilities from cumulative claims development triangles.

Let accident years be indexed by $a = 1, \dots, A$
and development periods be indexed by $j = 0, \dots, J$.

Let $C_{a,j}$denote the cumulative claims for accident year $a$observed at development period j. Let $C_{a,j}$ denote the cumulative claims for accident year a observed at development period j. For clarity, the reserving time structure is defined as follows. The index α denotes the accident year and j denotes the development period measured in months after occurrence. The maximum observed development period is denoted by J. The valuation development period is denoted by j_v , representing the time at which reserves are calculated based on observed claims data, while the ultimate development period is denoted by j_u, representing the development point at which claims are considered fully developed. In this study, $j_v = 12$ months and $j_u = 36$ months.

The cumulative claim amount C_{a,j_v} therefore represents the observed cumulative claims at valuation, and the incurred but not reported (IBNR) reserve corresponds to the difference between the estimated ultimate claims and the cumulative claims at the valuation period.

1. Standard CL Formulation

Under equally spaced development periods, the age-to-age development factors are defined as

$$\widehat{f_j} = \frac{\sum_{a=1}^{A-j} C_{a,j+1}}{\sum_{a=1}^{A-j} C_{a,j}}, \quad j = 0, \dots, J-1 \qquad (3)$$

Projected cumulative claims are obtained recursively as

$$\hat{C}_{a,j+1} = \hat{C}_{a,j} \cdot \hat{f}_j \qquad (4)$$

The CL method assumes proportional development stability across accident years, independence between accident years, and that historical development patterns are predictive of future development.

2. Development Structure in This Study

In this study, cumulative claims are observed at discrete and irregular development months given by

$$j \in \{0,3,6,12,24,36\} \qquad (5)$$

Because development intervals are irregular, let j^+denote the next observed development period following j.

The development factors are therefore defined as

$$\hat{f}_j = \frac{\sum_{a=1}^{A-j} C_{a,j^+}}{\sum_{a=1}^{A-j} C_{a,j}} \qquad (6)$$

Projected cumulative claims become

$$\hat{C}_{a,j^+} = \hat{C}_{a,j} \cdot \hat{f}_j \qquad (7)$$

This formulation generalizes the standard CL estimator to irregular development intervals while preserving the proportional development assumption.

3. IBNR Estimation

Let the valuation development period be denoted by j_v and the ultimate development period by j_u.

In this study,

$$j_v = 12, j_u = 36 \qquad (8)$$

For each accident year a, the CL estimate of IBNR at valuation date j_v is defined as

$$\hat{R}_a^{CL} = \hat{C}_{a,j_u} - C_{a,j_v} \qquad (9)$$

Thus, the study-specific IBNR estimate becomes

$$\hat{R}_a^{CL} = \hat{C}_{a,36} - C_{a,12} \qquad (10)$$

H. XGBoost Model

XGBoost is employed as a stand-alone ML model to predict incremental claims. The model constructs an additive ensemble of regression trees. Consistent with modern micro-level reserving approaches, the model is applied at the individual claim level, allowing claim characteristics and policy attributes to influence the prediction of future incremental payments.

The predicted incremental claim is given by

$$\hat{y}_{a,j} = \sum_{t=1}^{T} f_t\left(x_{a,j}\right) \qquad (11)$$

Model parameters are estimated by minimizing the regularized objective function

$$L(\phi) = \sum_{a=1}^{A}\sum_{j=0}^{J} l\left(\Delta C_{a,j}, \hat{y}a,j\right) + \sum t = 1^T \Omega(f_t) \qquad (12)$$

1. Aggregation to Cumulative Claims

Predicted cumulative claims are obtained as

$$\hat{C}_{a,j}^{XGB} = \sum_{j'=0}^{j} \hat{y}_{a,j'} \qquad (13)$$

The predicted ultimate claim for accident year $a$is

$$\hat{C}_{a,J}^{XGB} = \sum_{j=0}^{J} \hat{y}_{a,j} \qquad (14)$$

The predicted IBNR reserve at valuation development period $j_v = 12$ is

$$\hat{R}_a^{XGB} = \hat{C}_{a,J}^{XGB} - C_{a,12}^{obs} \qquad (15)$$

2. Hybrid Framework (Proposed)

To combine the strengths of both approaches, a hybrid framework is introduced.

The hybrid claim-level prediction for each claim i is defined as

$$\hat{y}_i^{Hybrid} = \begin{cases} \hat{y}_i^{CL}, & y_i \le \tau \\ \hat{y}_i^{XGB}, & y_i > \tau \end{cases} \quad (16)$$

XGBoost incremental predictions are aggregated by accident year and development period

$$\hat{Y}_{a,j}^{XGB} = \sum_{i \in C_{a,j}^{XGB}} \hat{y}_{i,j}^{XGB} \quad (17)$$

Cumulative predictions are obtained as

$$\hat{Y}_{a,j}^{XGB,cum} = \sum_{j'=0}^{j} \hat{Y}_{a,j'}^{XGB} \quad (18)$$

Finally, the hybrid IBNR reserve is computed as

$$\hat{R}_{Hybrid} = \sum_{i \in all\ claims} \hat{y}_i^{Hybrid} = \sum_{i \in small\ claims} \hat{y}_i^{CL} + \sum_{i \in large\ claims} \hat{y}_i^{XGB} \quad (19)$$

The CL component operates strictly at the aggregate accident-year development triangle level, consistent with actuarial reserving practice. No accident-year projections are allocated to individual claims, and no proportional or exposure-based redistribution of CL estimates to claim records is performed. Instead, XGBoost predicts future incremental payments beyond the 12-month valuation age at the individual claim level. These predicted incrementals are aggregated by accident year and development period to construct projected cumulative development triangles. The CL model is applied only to the small-claim segment triangle, while the aggregated XGBoost predictions represent the large-claim segment. The two segment triangles are modeled independently and then recombined by summing their cumulative projections to obtain the portfolio-level ultimate losses. The IBNR reserve is computed as the difference between the predicted ultimate cumulative losses at 36 months and the observed cumulative payments at the 12-month valuation point. This ensures methodological comparability while avoiding artificial allocation of aggregate CL projections to individual claims and prevents double counting between model components.

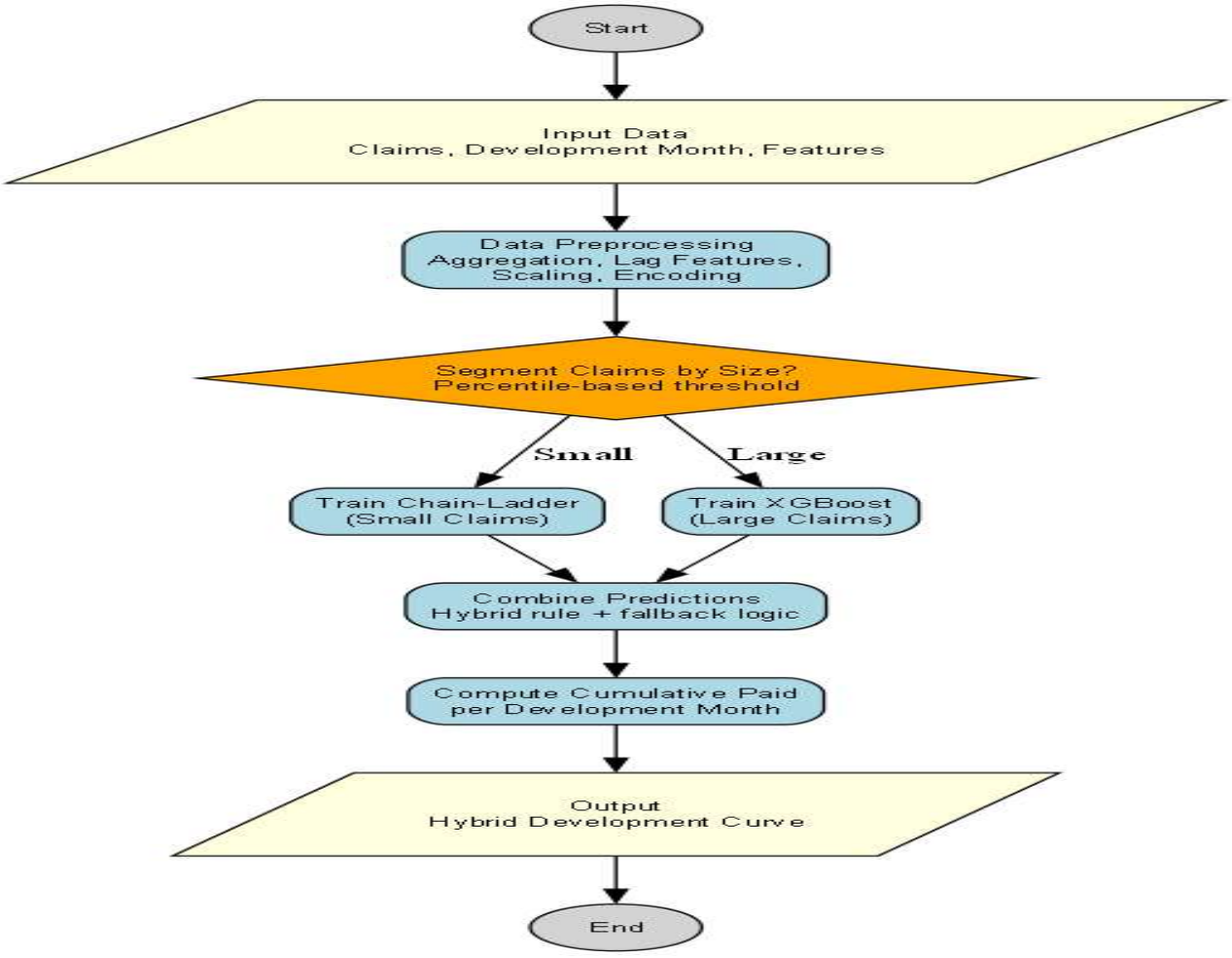


Figure 2 Hybrid CL–XGBoost workflow with percentile-based claim amount segmentation.

3. Uniform Hybrid Model for Segmentation Evaluation

To determine whether improvements in predictive accuracy arise from hybridization alone or from the proposed segmentation strategy, an additional benchmark model was constructed. This model represents a uniform (non-segmented) hybrid framework in which both the CL and XGBoost models are trained on the full dataset without claim segmentation.

Because both models generate predictions for the same future development periods, their outputs are combined using a convex weighting scheme:

$$\widehat{IBNR}_{UH} = \alpha \widehat{IBNR}_{CL} + (1-\alpha)\,\widehat{IBNR}_{XGB} \quad (20)$$

where

- $\widehat{IBNR}_{UH}$ denotes the reserve predicted by the uniform hybrid model,
- $\widehat{IBNR}_{CL}$ represents the reserve estimated by the CL method, and
- $\widehat{IBNR}_{XGB}$ denotes the reserve predicted by the XGBoost model.

The parameter $\alpha \in [0,1]$ controls the relative contribution of the actuarial and ML components.

The mixing weight α was selected using the validation accident year (2017) by minimizing the validation Mean Absolute Error (MAE). The optimal value obtained from the validation experiment was $\alpha = 0.96$. This procedure ensures that model tuning is performed without using information from the test accident years, thereby preventing bias in the out-of-sample evaluation.

The uniform hybrid model, therefore, serves as a benchmark for evaluating whether the performance improvements observed in the proposed model are attributable to hybridization itself or to the claim segmentation strategy introduced in this study.

4 Neural Network-Chain Ladder

The Neural Network-Chain Ladder (NNCL) model is used as a benchmark reserving approach that extends the classical CL method by allowing development factors to depend on observable claim characteristics through a neural-network structure (Wüthrich, 2018). Unlike the proposed hybrid framework, NNCL is applied to the full dataset without claim segmentation and therefore evaluates whether segmentation, rather than neural-network learning alone, contributes to improved predictive performance.

(i) Classical CL Model

In the case of the classical CL model, let us consider the random variable $C_{i,j}$ denoting the cumulative claims for accident year ($i = 1,\dots,I$) and development period ($j = 0,\dots,J$). The expected value of $C_{i,j}$, conditional to the information up to the previous development period, can be expressed as follows:

$$E\left[C_{i,j}\middle|F_{i+j-1}\right] = f_{j-1}C_{i,j-1}, \quad (21)$$

where F_{i+j-1} denotes the information set available up to development period $j-1$, including all observed cumulative claims up to calendar time $i+j-1$, where $f_{j-1} > 0$ is a development factor that depends solely on the development period.

From the above equation we can see that one of the assumptions of the classical CL model is that of the homogeneity of the portfolio, since the same development factor applies to all accident years and types of claims.

(ii) Decomposition of Cumulative Claims Using Features

The feature vector $x \in X$ represents observable characteristics associated with each claim observation. The set X denotes the collection of all possible observable claim profiles defined by combinations of claim-level and policy-level covariates present in the dataset.The reserving analysis is performed at the individual claim (micro) level, after which predicted developments are aggregated to obtain accident-year IBNR reserve estimates. The features are obtained from the claim-level insurance dataset and include both claim-specific and policy-related variables. Claim-level variables describe individual claim behavior, such as reporting delay (Reporting_Delay_Months), development period (Development_Month), and accident year (Accident_Year). Policy and risk-related variables describe exposure and risk characteristics of the insured entity, including premium, type of risk, geographical area, and vehicle-related attributes (e.g., vehicle value and engine power). These covariates are used as inputs to the neural network so that development factors can vary across heterogeneous claim profiles rather than remaining constant across the entire portfolio.

Categorical variables are numerically encoded prior to training, and all features are standardized to ensure numerical stability during neural-network optimization. Conditioning the development factors on these observable covariates relaxes the homogeneity assumption of the classical CL method and allows the model to account for heterogeneity in claim development behavior across different claim and policy characteristics. We can write the cumulative claims as follows:

$$C_{i,j} = \sum_{x \in X} C_{i,j}(x) \quad (22)$$

Equation (19) represents an aggregation across claims sharing the same observable feature profile, so that cumulative claims at the accident-year level are obtained by summing the contributions of individual claim profiles. Given that the cumulative claims related to a certain feature profile (x) have a non-zero exposure $\left(C_{i,j-1}(x) > 0\right)$, the Neural Network-Chain Ladder (NNCL) model assumes that the expected value of the cumulative claims, conditional to the information available up to the previous development period, can be expressed as follows:

$$E\left[C_{i,j}(x)\middle|F_{i+j-1}\right] = f_{j-1}(x)C_{i,j-1}(x) \quad (23)$$

where $f_{j-1}(x)$ represents the development factor dependent on the feature profile. If $f_{j-1}(x) \equiv f_{j-1}$ for every (x), we recover the classical chain ladder model in (18).

(iii) Representation of Development Factors using Feedforward Neural Networks

To ensure that the development factors remain greater than or equal to zero, the NNCL model adopts a representation based on a log-link:

$$f_{j-1} = exp\left(g_{j-1}(x)\right) \quad (24)$$

where $g_{j-1}(x)$ represents an unknown nonlinear function that can be approximated by a feed-forward neural network.

A separate neural network is developed for each development period $(j = 1, \ldots, J)$ by using the complete set of available claims. Each neural network has two hidden layers. Two hidden layers were selected to balance flexibility and generalization. A single hidden layer produced underfitting and failed to capture heterogeneous development behavior, while deeper networks increased model variance and overfitting due to limited observations per development period. The number of neurons was selected using validation-set error minimization. The centered sigmoid activation function was adopted because it preserves non-negativity of development factors under the log-link formulation and improves numerical stability during training.

Hidden Layer First

Let $\left(x = (x_1, \ldots, x_d)\right)$ represent the (d)-dimensional vector of features.

$$z_k^{(1)}(x) = \varphi\left(w_{k,0}^{(1)} + \sum_{l=1}^{d} w_{k,l}^{(1)} x_l\right), k = 1, \ldots, q_1 \quad (25)$$

with ($\varphi(\cdot)$) representing the centered sigmoid function.

Second Hidden Layer

Similarly, we obtain

$$z_k^{(2)}(x) = \varphi\left(w_{k,0}^{(2)} + \sum_{l=1}^{q_1} w_{k,l}^{(2)} z_l^{(1)}(x)\right), \quad k = 1, \ldots, q_2 \quad (26)$$

The activation function $\left(\varphi(\cdot)\right)$ is the centered sigmoid, defined as follows:

$$\varphi(u) = \frac{2}{1 + e^{-u}} - 1 \quad (27)$$

that maps the input values to the range$\left((-1,1)\right)$ and helps improve the stability of the computations during the training process.

Layer Output

The network output is given by

$$g_{j-1}(x) = \beta_0 + \sum_{k=1}^{q_2} \beta_k z_k^{(2)}(x) \quad (28)$$

and it completely determines the feature-dependent development factor $\left(f_{j-1}(x)\right)$ through (21).

(iv) Estimation of the Parameters of the Model

The parameters of the model are estimated separately for each development period by minimizing a weighted least squares loss function, consistent with the variance structure proposed by Mack:

$$L_j = \sum_{i=1}^{I-j} \sum_{x: C_{i,j-1}(x) > 0} C_{i,j-1}(x) \left(\frac{C_{i,j}(x)}{C_{i,j-1}(x)} - f_{j-1}(x)\right)^2 \quad (29)$$

The optimization is carried out by means of a gradient descent algorithm with backpropagation.

The number of hidden units (q_1, q_2) is chosen using out-of-sample validation to find the best compromise between the level of detail of the predictions and the level of complexity of the model.

(v) Cells Without Exposure

For the cells of the feature matrix that do not present any cumulative exposure, i.e., $\left(C_{i,j-1}(x) = 0\right)$, the recursive relationship in (20) cannot be applied. Therefore, these cells are aggregated and projected onto a common chain-ladder type structure:

$$C_{i,J}^{*} = C_{i,I-i} \prod_{j=I-i}^{J-1} g_j^{(i)} \quad (30)$$

where $\left(g_j^{(i)}\right)$ are the development factors estimated from the aggregated historical data associated with the cells of the feature matrix without exposure.

3.5.6 Claims and Reserving Estimates Ultimate

The estimated ultimate claims for accident year (i) are given by

$$\hat{C}_{i,J} = \sum_{x \in X} C_{i,I-i}(x) \prod_{j=I-i}^{J-1} \hat{f}_j(x) + C_{i,J}^{*} \quad (31)$$

The corresponding incurred but not reported (IBNR) reserve is

$$\hat{R}^{NNCL} = \hat{C}_{i,J} - C_{i,I-i} \quad (32)$$

(vi) Evaluation of Models

The performance of the model is evaluated using cumulative claims estimates at the accident-year level to allow for direct comparison with traditional actuarial methods and to ensure consistency with machine learning approaches. The evaluation of cumulative amounts directly represents the quantities of interest in actuarial reserving, as reserve adequacy and stability are typically assessed at the aggregate accident-year level rather than at intermediate levels of incremental predictions. This evaluation framework preserves the internal coherence of the run-off triangle and provides a valid basis for comparing models with fundamentally different structures.

Predictive accuracy is assessed using three standard performance metrics: Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and Mean Absolute Percentage Error (MAPE). Collectively, these measures capture the average magnitude of prediction errors (MAE), the sensitivity to large deviations from predicted values (RMSE), and the relative accuracy of each model's predictions (MAPE), thereby providing a comprehensive assessment of model performance.

IV. RESULTS AND DISCUSSION

This section presents and interprets the empirical performance of the proposed segmentation-based hybrid reserving framework. The predictive accuracy of the Hybrid model is compared with the CL, standalone XGBoost, and Neural Network–Chain Ladder benchmark models using standard error metrics and graphical evaluation. The objective

is to assess whether segmentation improves IBNR prediction accuracy under realistic reserving conditions.

A. Comparative Performance of Models

This subsection evaluates the predictive performance of the Hybrid model relative to the CL, XGBoost, and Neural Network–Chain Ladder models using Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE).

Figure 3 presents the observed cumulative IBNR for the test accident years together with the model predictions. All models capture the upward development trend from accident year 2018 to 2019. However, the Hybrid model produces predictions closest to the observed cumulative IBNR values in both years. The Neural Network–Chain Ladder model follows closely, while the Uniform Hybrid model shows moderate improvement over the standalone approaches. In contrast, the CL and standalone XGBoost models exhibit larger deviations from the observed claim development.

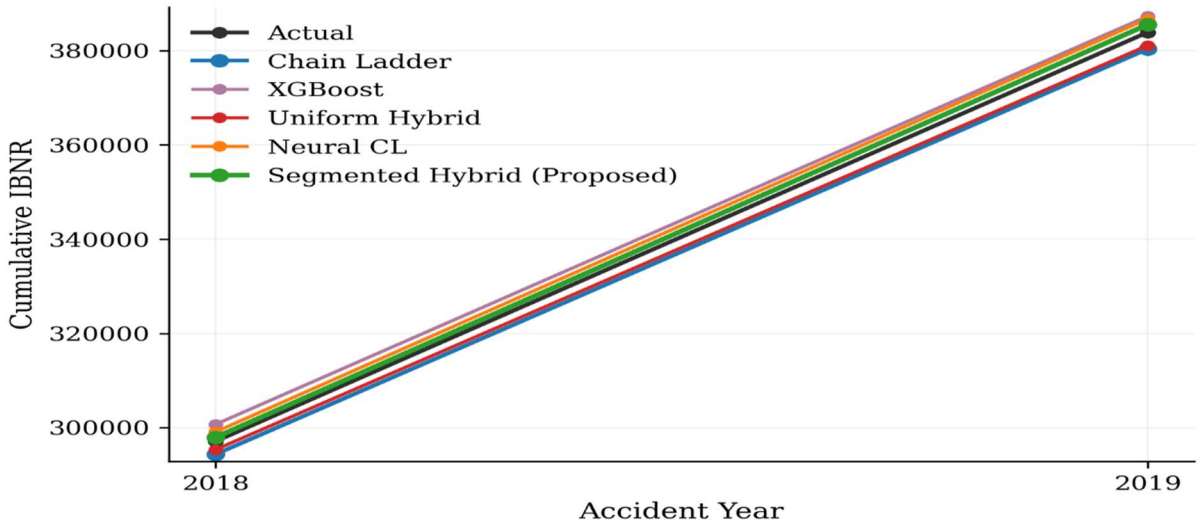


Figure 3 Observed cumulative IBNR claims compared to the predicted cumulative IBNR claims for the test years.

The predictive accuracy of the reserving models is summarized in Table 2. The classical CL and standalone XGBoost models exhibit relatively high prediction errors when applied to the heterogeneous portfolio. The Neural Network–Chain Ladder model improves upon the classical CL but still produces larger errors than the hybrid approaches.

Introducing a uniform hybrid framework that combines CL and XGBoost without segmentation reduces prediction errors moderately (MAE = 1,401.71). However, the proposed segmentation-based hybrid model produces the lowest error across all evaluation metrics, with MAE = 819.28 and RMSE = 819.39.

Relative to the classical CL baseline, the segmented hybrid model reduces MAE by approximately 54%, while relative to the uniform hybrid model the error decreases by approximately 41.6%. This indicates that the performance improvement arises primarily from heterogeneity-aware claim segmentation rather than hybridization alone.

Table 2 Comparative Performance of Reserving Models

Model	MAE	RMSE	MAPE (%)
CL	1,774.51	2,026.08	0.92
XGBoost	1,892.93	2,549.73	0.71
Neural Network–Chain Ladder	1,427.56	1,537.80	0.83
Uniform Hybrid (CL + XGBoost, no segmentation)	1,401.71	1,447.68	0.90
Segmented Hybrid (Proposed Model)	819.28	819.39	0.60

Figure 4 shows the signed prediction errors for each model by accident year. The Hybrid model produces errors closest to zero across both years, demonstrating improved stability of reserve estimates. The CL model underestimates IBNR in 2018 and continues to produce negative prediction errors in 2019, while the standalone XGBoost model displays inconsistent prediction direction across accident years, substantially overestimating reserves in 2018 but slightly underestimating them in 2019. The Uniform Hybrid model reduces the magnitude of the prediction errors relative to the standalone models, but still exhibits noticeable underestimation in both years. In contrast, the Neural Network–Chain Ladder model produces positive errors in both years. Overall, the improved stability of the Hybrid model indicates that segmentation mitigates the influence of large, volatile claims on development estimation.

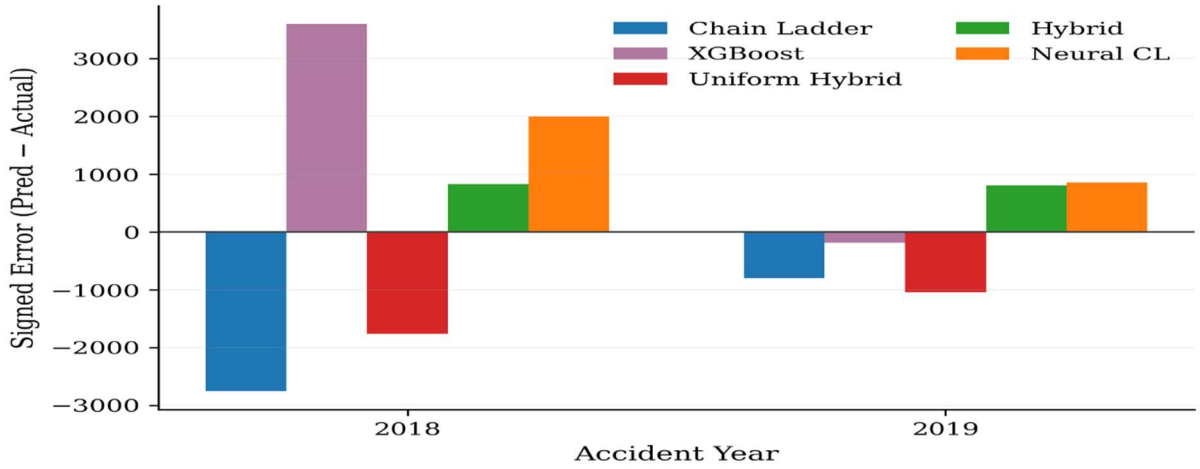


Figure 4 IBNR prediction errors for each of the five models for the test accident Years.

B. Sensitivity Analysis of the Segmentation Threshold

To justify the selection of the segmentation threshold, a sensitivity analysis was conducted using the validation dataset. Candidate percentile thresholds were determined from the training data and applied to the validation accident year without recalibration. The hybrid model was evaluated using validation Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) computed on predicted IBNR reserves.

Table 3 Validation performance of the hybrid reserving model across candidate claim segmentation thresholds used to determine the optimal 65th percentile cutoff.

Percentile	Threshold Value (€)	Validation MAE	Validation RMSE
60th	21.45	1842.35	1910.82
63rd	24.86	921.74	948.63
65th	27.72	510.20	533.41
67th	30.89	648.91	676.52
70th	36.14	1276.48	1318.94

As shown in Table 3**,** the 65th percentile threshold minimizes both MAE and RMSE on the validation set. As expected, RMSE values are slightly higher than MAE across all thresholds due to the stronger penalization of larger deviations. Lower thresholds (60th–63rd percentiles) insufficiently separate large claims from typical claims, leading to distortion in development factors within the CL component. Conversely, higher thresholds (67th–70th percentiles) over-segment the portfolio, reducing the amount of data available to train the machine learning component and increasing prediction variability.

The 65th percentile therefore represents a balance between actuarial stability and machine learning flexibility. Based on this out-of-sample validation performance, the 65th percentile segmentation threshold was selected as the operational threshold for the hybrid reserving framework.

C. Robustness to Reporting Delay Variability

To evaluate whether the performance of the proposed framework depends on a particular reporting delay assumption, the reserving models were re-estimated under alternative variance specifications of the reporting delay distribution. Lower-variance and higher-variance delay scenarios were generated while maintaining the same mean delay structure. The normal-variance case corresponds to the baseline results presented in Section IV-A, table 2. Table 4 reports prediction errors under the three scenarios. Across all variance settings, the Hybrid model consistently produces the lowest MAE, RMSE, and MAPE. Although prediction errors increase as variance increases, the relative ranking of models remains unchanged. This indicates that the performance improvement of the hybrid framework is not dependent on a single delay specification. The results therefore demonstrate that the

proposed segmentation-based hybrid model is stable under different reporting delay structures.

Table 4 IBNR prediction errors under different variance scenarios

Variance Scenario	Model	MAE	RMSE	MAPE (%)
Lower Variance	CL	1,521.34	1,642.87	0.81
	XGBoost	1,602.18	1,934.62	0.63
	Neural Network–Chain Ladder	1,219.45	1,305.72	0.74
	Uniform Hybrid	1,524.57	1,654.54	0.80
	Hybrid Model	693.52	701.88	0.54
Normal Variance	CL	1,774.51	2,026.08	0.92
	XGBoost	1,892.93	2,549.73	0.71
	Neural Network–Chain Ladder	1,427.56	1,537.80	0.83
	Uniform Hybrid	1,401.71	1,447.68	0.90
	Hybrid Model	819.28	819.39	0.60
Higher Variance	CL	2,184.73	2,842.51	1.15
	XGBoost	2,476.90	3,684.32	0.94
	Neural Network–Chain Ladder	1,782.66	2,241.58	1.02
	Uniform Hybrid	2,196.42	2,876.18	1.14
	Hybrid Model	1,094.81	1,248.63	0.78

D. Performance of the proposed Hybrid Framework

This subsection evaluates the predictive behavior of each component of the Hybrid framework within its assigned claim segment. Figure 5 illustrates cumulative IBNR predictions by segment. The CL component accurately predicts the small-claim segment across both accident years, reflecting stable development patterns. In contrast, the XGBoost component better captures the development of the large-claim segment, particularly in accident year 2019 where nonlinear claim behavior is more pronounced and exhibits higher variability. This indicates that the machine-learning component adapts more effectively to irregular and high-severity claim dynamics than the actuarial development-factor approach.

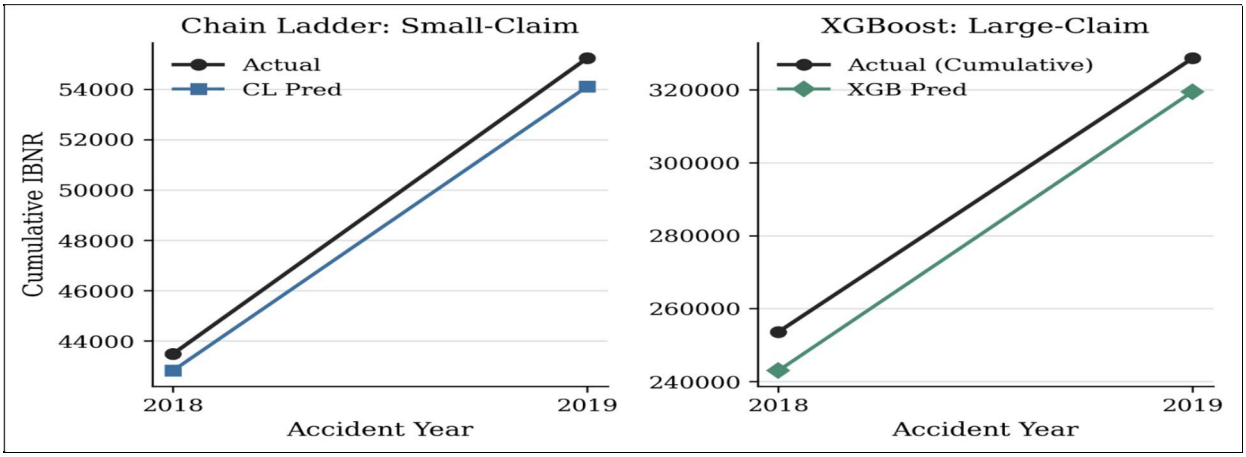


Figure 5 Cumulative IBNR estimates by the components of the Hybrid model for test accident years by segments

Table 5 presents the segment-specific prediction errors. The CL method achieves very low absolute percentage errors for the small-claim segment, confirming that traditional development factors perform well under homogeneous conditions. Conversely, XGBoost produces improved predictions when applied only to the large-claim segment, where greater variability and nonlinear relationships exist.

These findings demonstrate that segmentation reduces heterogeneity within each modeling subset. When applied to the full dataset, the CL method is sensitive to extreme claims, leading to distorted development factors. Similarly, XGBoost applied to all claims encounters high variance due to numerous small or zero incremental payments. Restricting each method to the segment best suited to its assumptions improves overall predictive performance.

The results therefore reflect the bias–variance trade-off between actuarial and ML models. CL provides stable low-variance estimates for homogeneous small claims, while XGBoost captures complex nonlinear behavior for large claims. The Hybrid framework leverages both properties by allocating each method to the segment where it performs best.

Table 5 Prediction errors for the hybrid components for each segment

Model	Segment	2018 Error	2018 APE (%)	2019 Error	2019 APE (%)	Mean Error (2018–2019)	Mean APE (%)
CL	Small-claim IBNR	136.06	0.05	114.70	0.13	125.38	0.09
XGBoost	Large-claim IBNR	696.78	0.23	691.02	0.80	693.90	0.52

E. Discussion

The empirical results demonstrate clear differences in predictive behavior among the actuarial, machine-learning, and hybrid reserving methods. The classical CL method performs adequately for stable and homogeneous claim development but deteriorates when large claims distort development factors, as reflected by its higher MAE and RMSE values across all variance scenarios. In contrast, the standalone XGBoost model captures nonlinear claim patterns but exhibits higher variability in prediction accuracy when applied to the full heterogeneous dataset.

The segmentation-based Hybrid framework improves performance by assigning each method to the segment most consistent with its modeling assumptions. The CL component produces accurate predictions in the small-claim segment, while XGBoost better models the volatile large-claim segment. This complementary allocation stabilizes development factor estimation and reduces prediction error, which is evident from the consistently lower error measures reported in Table 2.

The sensitivity analysis around the segmentation threshold further supports this conclusion. Validation results (Table 3) show that the 65th percentile minimizes both MAE and RMSE, indicating that the selected threshold balances development stability and learning flexibility. Lower thresholds fail to adequately isolate large claims, whereas higher thresholds reduce the amount of information available for machine learning, leading to increased prediction variance.

Robustness analysis under alternative reporting delay variance assumptions (Table 4) shows that although prediction errors increase as delay variability increases, the relative ranking of models remains unchanged. The Hybrid model consistently produces the lowest prediction errors across lower, normal, and higher variance scenarios, indicating that the performance improvement is not dependent on a single delay specification.

These findings indicate that hybrid reserving benefits from addressing claim heterogeneity rather than replacing actuarial methods entirely. The framework preserves actuarial interpretability in the stable claim segment while applying machine learning selectively where nonlinear behavior is present.

The proposed model, however, produces point reserve estimates and does not quantify reserve uncertainty. Additionally, performance may depend on the segmentation threshold and historical claim availability. Future research may therefore extend the framework to stochastic reserve estimation and validation across multiple insurance portfolios.

V. CONCLUSION

This study proposed a segmentation-based hybrid reserving framework that combines the traditional CL method with the Extreme Gradient Boosting (XGBoost) algorithm. The model was applied to claim size segmentation and evaluated in a valuation setting in which cumulative claims observed at 12 months were projected to an ultimate development horizon of 36 months using out-of-sample accident years.

The empirical results show that the Hybrid model consistently outperforms the classical CL, Neural Network–Chain Ladder, and standalone XGBoost models across all evaluation metrics. The Hybrid framework achieves the lowest Mean Absolute Error (MAE), Root Mean Square Error (RMSE), and Mean Absolute Percentage Error (MAPE), and its predicted cumulative IBNR closely follows observed claim development patterns. The performance improvement arises from the complementary strengths of the two components: the stability and interpretability of the CL method for small and homogeneous claims, and the flexibility of XGBoost in capturing nonlinear dynamics in large and heterogeneous claims.

The findings indicate that reserving accuracy depends strongly on claim heterogeneity. Allocating models according to claim characteristics improves predictive performance by stabilizing development factor estimation while allowing flexible learning where complex behavior

exists. The results therefore support the use of hybrid actuarial–ML frameworks as a practical approach to operational reserving when combined with segmentation and out-of-sample validation.

This study has several limitations. The framework produces point estimates of reserves and does not quantify reserve uncertainty. In addition, accident and reporting dates were synthetically generated to construct development patterns, and model performance may vary for portfolios with different reporting structures or claim histories. Future research may extend the approach to stochastic reserve estimation, alternative segmentation strategies, and validation across multiple insurance portfolios.

Overall, the study provides empirical evidence that segmentation-based hybrid reserving can improve predictive accuracy while preserving actuarial interpretability in non-life insurance reserving.

AUTHOR CONTRIBUTIONS

Nhial Mabior: conceived the study, designed the methodology, performed data analysis, implemented the models, and drafted the manuscript.

Oscar Ngesa: supervised the research, contributed to methodological refinement, and reviewed the manuscript. Jane Aduda: provided academic guidance, assisted with interpretation of actuarial results, and reviewed the manuscript.

ACKNOWLEDGEMENT

The authors acknowledge the academic environment and support provided by their respective institutions, which facilitated the completion of this research. The authors also thank colleagues for constructive discussions that helped improve the clarity of the manuscript.

REFERENCES

Antonio, K. and Plat, R. (2021). Statistical learning applied to claims reserving. *Risks*, 9(2), 28.

Avanzi, B., Li, Y., Wong, B. and Xian, A. (2023). Modern machine learning approaches to insurance reserving. *Insurance: Mathematics and Economics*, 110, 1–17.

Avanzi, B., Li, Y., Wong, B. and Xian, A. (2024). Ensemble distributional forecasting for insurance claims. *Insurance: Mathematics and Economics*, 114, 52–70.

Blier-Wong, C., Cossette, H. and Marceau, E. (2020). Machine learning in actuarial science: A review of applications in claims reserving. *Risks*, 8(4), 117.

Bücher, A. and Rosenstock, M. (2023). Neural networks for loss reserving. *European Actuarial Journal*, 13, 245–272.

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

Dong, Y. and Quan, Z. (2025). Interpretability and risk assessment of machine learning reserving models. *Annals of Actuarial Science*, 19(1), 1–25.

Feng, R. and Li, Y. (2023). Robust chain ladder methods for loss reserving with extreme claims. *ASTIN Bulletin*, 53(2), 421–447.

Henckaerts, R., Antonio, K., Verbelen, R. and Dhaene, J. (2021). A review of loss reserving methods. *European Actuarial Journal*, 11, 1–34.

Hiabu, M. and Müller, M. (2021). Micro-level reserving and machine learning in non-life insurance. *Insurance: Mathematics and Economics*, 97, 114–129.

Hiabu, M., Nielsen, J. and Verrall, R. (2023). Machine learning and insurance reserving: Opportunities and challenges. *Scandinavian Actuarial Journal*, 2023(5), 395–418.

Krisdiantha, S., Purwono, R. and Sari, N. (2023). Chain ladder reserving under heterogeneous portfolios. *Journal of Risk and Financial Management*, 16(3), 145.

Kuo, K. (2019). DeepTriangle: A deep learning approach to loss reserving. *Risks*, 7(3), 97.

Laub, P., Taimre, T. and Pollett, P. (2025). Bias-variance trade-off in machine learning actuarial models. *Insurance: Mathematics and Economics*, 120, 65–82.

Mahohoho, T., Thupeng, W. and Sigauke, C. (2023). Artificial intelligence in actuarial reserving. *Risks*, 11(6), 110.

Manathunga, C. and Zhu, X. (2022). Machine learning performance stability in insurance data. *Journal of Computational and Applied Mathematics*, 410, 114208.

Mazzi, L., Savelli, N. and Clemente, G. (2024). Heterogeneity in insurance loss development. *ASTIN Bulletin*, 54(1), 101–128.

Poufinas, T., Galanos, G. and Papadopoulos, A. (2023). Predictive analytics in insurance risk management. *Journal of Risk Finance*, 24(2), 180–198.

Rahul, A. (2020). Gradient boosting techniques for claims reserving. *Journal of Risk and Financial Management*, 13(11), 275.

Saida, A., Rahman, M. and Karim, S. (2023). Cluster-based segmentation for insurance reserving. *Journal of Insurance and Financial Management*, 6(2), 45–60.

Saoudi, A., Benali, M. and Boussaada, I. (2023). Evaluation of classical reserving models in non-life insurance. *Journal of Applied Statistics*, 50(7), 1405–1422.

Song, J. and Heo, J. (2022). Predictive modelling of claim severity using machine learning. *Expert Systems with Applications*, 197, 116641.

Suwardi, A. and Purwono, R. (2021). Improved chain ladder reserving methods under outlier conditions. *Journal of Physics: Conference Series*, 1725, 012045.

Wüthrich, M. (2018). Neural network embedding of the chain ladder method. *European Actuarial Journal*, 8(2), 409–431.

England, P. D., and Verrall, R. J. (2002). Stochastic claims reserving in general insurance. British Actuarial Journal, 8(3), 443–544.

Wüthrich, M. V., and Merz, M. (2008). Stochastic Claims Reserving Methods in Insurance. Wiley.

APPENDIX

(a) (b)

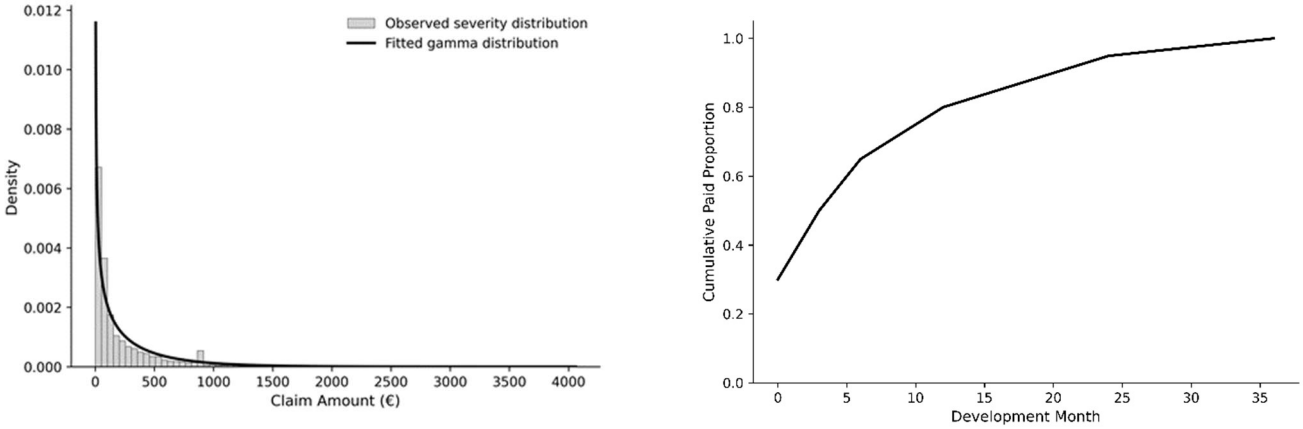


Figure A1 Validation of claim severity distribution and development completeness.

(a) Observed claim severity histogram with fitted Gamma distribution, showing a heavy right-tailed payment structure consistent with insurance loss severity modeling.
(b) Empirical cumulative paid development curve by development month, illustrating that claim payments approach stability near 36 months, supporting the selection f 36 months as the operational ultimate development period

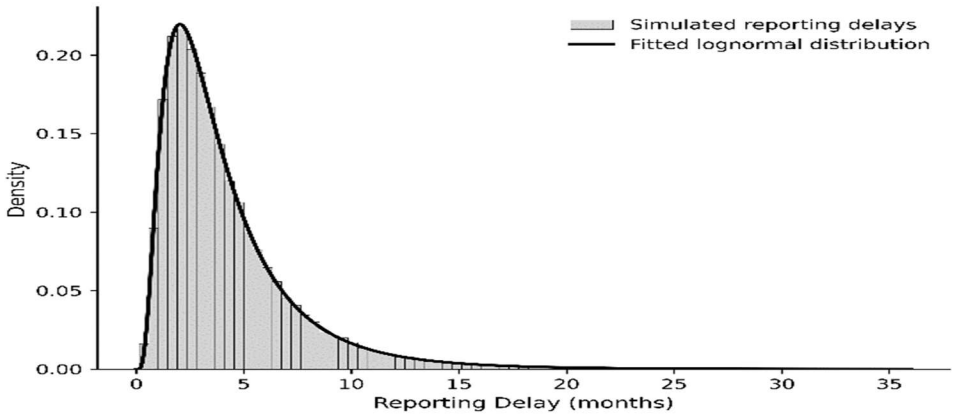


Figure A2 Validation of reporting delay distribution assumption.

Histogram of simulated reporting delays with fitted Lognormal distribution. The close agreement between observed and fitted densities indicates that the Lognormal distribution adequately represents reporting delay behavior, justifying its use in the simulation of accident and reporting dates.