Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10, Number 1, 2026
DOI: 10.30829/zero.v10i1.28134
Page: 168-179

Hierarchical Ensemble Actuarial Method for Motor Claim Reserving under Indonesia's PSAKBI

[1] Mulawarman Awaloedin
Actuarial Studies Program, STMA Trisakti, Indonesia.

Article Info

Article history:

Accepted 25 March 2026

Keywords:

Actuarial Reserving;
Claim Reserves;
Ensemble Learning;
Hybrid Model;
Indonesia Case Study;
Motor Insurance;
PSAKBI.

ABSTRACT

This study develops and evaluates a hierarchical ensemble actuarial approach for motor claim reserving under Indonesia's PSAKBI framework. The method integrates traditional actuarial techniques with modern machine learning models in a structured ensemble design to enhance predictive accuracy, reliability, and transparency. Using motor insurance claim data, the ensemble was compared against conventional single-model reserving practices. Results show that the proposed approach achieves lower prediction error (MSE = 220.3), accurate calibration (94.7% coverage), and more stable reserve estimates across accident years. Beyond statistical performance, the design emphasizes interpretability by tracing predictions to weighted contributions of base models, thereby avoiding black-box behavior. These findings highlight the practical relevance of hybrid ensemble reserving in regulated environments, offering a transparent and robust solution aligned with PSAKBI requirements. The study contributes to the literature by demonstrating how hybrid actuarial ensembles can balance methodological rigor, machine learning flexibility, and regulatory compliance in insurance reserving.

Corresponding Author:

Mulawarman Awaloedin,
Actuarial Studies Program,
STMA Trisakti, Jakarta, Indonesia.
Email: mulawarman@stma-trisakti.ac.id

1. INTRODUCTION

Over the past several decades, actuarial and insurance modeling has undergone a profound transformation, shifting from deterministic and aggregate-based approaches toward data-driven, probabilistic, and hybrid frameworks. Classical actuarial models such as the Generalized Linear Model (GLM) and Generalized Additive Model (GAM) have long served as foundational tools in tariff modeling, frequency–severity estimation, and claim reserving, both in life and non-life insurance [1], [2]. These models provided transparency and regulatory compliance, but their assumptions of linearity and parametric structure limited their ability to capture complex claim dynamics. As insurance portfolios became more diverse and data availability expanded, the demand for more granular and adaptive reserving methods grew. This evolution has led researchers and practitioners to increasingly adopt machine learning (ML) techniques, which offer flexibility and predictive power beyond the scope of traditional models [3]. In particular, Gradient Boosting has emerged as a well-established method for producing detailed estimates in individual loss reserving [4], while comparative studies between logistic regression and ML models such as XGBoost have highlighted both the superiority and trade-offs of newer approaches [5].

The rise of machine learning in actuarial science has been further accelerated by advances in computational capacity and the availability of large-scale insurance datasets. Methods such as Random Forest, Gradient Boosting Machines (GBM), and XGBoost have proven effective in capturing non-linear relationships among variables and achieving higher predictive accuracy compared to conventional models [6], [7]. At the same time, statistical approaches such as Zero-Inflated Poisson (ZIP) and Logistic Regression remain important for handling claim data with excess zeros, while Ridge Regression is applied to address multicollinearity [8], [9]. Deep learning architectures, particularly Long Short-Term Memory (LSTM) networks, have also been applied to model temporal claim development dynamics, offering new ways to capture sequential dependencies in claim payments [10]. More advanced approaches, including hierarchical reserving models and inverse probability weighting (IPW), attempt to overcome the limitations of aggregate models by leveraging granular claim-level information [11], [12]. At the same time, interpretability techniques such as SHAP (SHapley Additive Explanations) have been introduced to explain variable contributions in ML models, thereby addressing concerns about transparency and accountability [6]. These developments illustrate a paradigm shift in actuarial practice: from classical parametric approaches toward hybrid, data-driven models that emphasize both predictive accuracy and interpretability.

Despite these advances, several critical research gaps remain unresolved. First, there is no unified framework that combines the interpretability of traditional statistical models such as GLM and GAM with the flexibility and high accuracy of machine learning algorithms such as XGBoost and LSTM. Statistical models are easy to explain and comply with regulations but struggle to capture complex variable interactions. Conversely, machine learning models excel in predictive accuracy but often function as "black boxes," making them difficult for practitioners and regulators to understand [2]. Second, modeling claim granularity and dependency remains a challenge. Traditional loss triangle approaches tend to ignore correlations across lines of business, while models such as the Extended Deep Triangle only partially address this issue [13]. Third, current model performance evaluations predominantly emphasize accuracy-based metrics, such as AUC and Gini, which fail to adequately capture reliability, fairness, and transparency—dimensions that are critical under regulatory oversight [7], [8]. Moreover, efforts to leverage Generative Adversarial Networks (GANs) to generate synthetic data for stochastic forecasting are still limited, leaving reliability and interpretability largely untested.

Another unresolved issue is the difficulty of quantifying uncertainty in claim reserve estimation. While Bayesian Chain Ladder provides a principled framework for uncertainty quantification, it relies on strong prior assumptions and entails intensive computational requirements [14]. Interpretability challenges become even more pronounced when hybrid and stacking models are employed, as tracing the contribution of individual components to the final prediction is often non-trivial. Although models such as Explainable Boosting Machines (EBM) have begun to attract attention, their application within the insurance domain remains limited [2], [15]. Furthermore, claim reserve estimation remains difficult due to instability and randomness in claim data, requiring approaches that are not only accurate but also reliable in handling overdispersion and zero inflation. This situation creates an urgent need to focus not only on accuracy but also on accountability and trust in predictive systems.

To address these gaps, this study proposes a hybrid actuarial ensemble framework that integrates statistical rigor, machine learning flexibility, and explainable AI techniques. The framework emphasizes interpretability by tracing predictions to weighted contributions of base models, thereby avoiding black-box behavior. It also incorporates uncertainty quantification through Bayesian methods and bootstrap resampling, ensuring robustness in reserve estimation. By combining predictive accuracy, transparency, and regulatory compliance, the proposed framework aims to provide a practical solution for motor claim reserving under Indonesia's PSAKBI. This is particularly relevant in regulated environments, where reserving practices must satisfy both accuracy and transparency requirements. The novelty of this study lies in its unified hybrid ensemble framework that integrates statistical foundations, machine learning algorithms, and interpretability techniques into a single coherent methodology.

Filling these gaps is crucial to ensure that reserving and risk prediction models in the insurance industry can meet the challenges of data complexity, regulatory demands, and transparency requirements. Regulatory compliance requires models that can be clearly explained to stakeholders, while industry practice demands reliable and robust reserve estimates. Therefore, integrating statistical foundations with machine learning and explainable AI offers a pathway to models that are simultaneously accurate, transparent, and aligned with supervisory expectations. This study contributes to the literature by demonstrating how hybrid ensemble reserving can strengthen actuarial practice, improve prudential reserving, and support regulatory transparency in the Indonesian insurance industry. In addition, the study highlights the importance of scalability and adaptability, recognizing that models must be capable of handling larger datasets and more complex claim structures in diverse insurance contexts.

This study extends the foundational work of Taylor (2019) on granular reserving models and Blier-Wong et al. (2021) on machine learning applications in P&C insurance by demonstrating how statistical and ML methods can be integrated within a unified, interpretable ensemble framework specifically tailored for regulated environments. The proposed approach not only advances methodological innovation but also delivers tangible societal benefits by enhancing reserve accuracy, which directly contributes to insurer solvency and policyholder protection—a critical need in Indonesia's rapidly evolving insurance market.

The urgency of this research is further underscored by the evolving regulatory landscape in Indonesia. PSAKBI, as the governing framework for insurance reserving, imposes strict requirements on accuracy, prudence, and transparency. Insurers must not only produce reliable reserve estimates but also demonstrate that their methods are explainable and compliant with regulatory standards. In this context, the adoption of hybrid ensemble models represents a significant step forward, offering a balance between methodological sophistication and regulatory accountability. By explicitly linking methodological innovation to regulatory compliance, this study underscores the practical relevance of actuarial research in addressing industry challenges.

Failure to address these challenges promptly could exacerbate insurer insolvency risks and undermine consumer confidence in the insurance sector, making this research both timely and essential for financial stability

The proposed framework is summarized in Figure 1, which illustrates the integration of statistical foundations, machine learning flexibility, and interpretability techniques into a unified actuarial ensemble model. Taken together, these considerations establish a robust research pathway: from identifying industry challenges, to designing a hybrid actuarial ensemble framework, to validating its effectiveness through iterative experimentation. The result is not only a technically advanced solution but also one that is practical, transparent, and applicable within the real-world context of the Indonesian insurance industry. This study therefore represents a meaningful contribution to both academic literature and industry practice, bridging the gap between theoretical innovation and regulatory application.

By integrating statistical foundations, machine learning algorithms, and explainable AI techniques into a single coherent framework, this study advances the frontier of hybrid actuarial modeling—offering a template that is both methodologically innovative and practically relevant for regulated insurance markets like Indonesia.

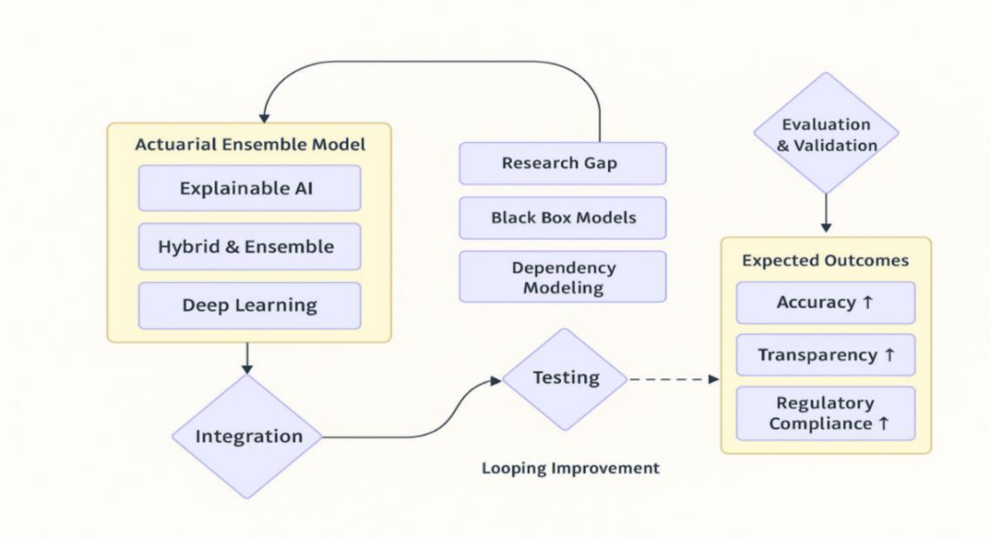

Figure 1. A Framework for an Explainable Actuarial Ensemble Model

2. RESEARCH METHOD

2.1 Research Design

The study adopts a quantitative analytical–computational design to develop and evaluate a hybrid reserving framework that integrates actuarial and machine learning methods. The objective is to compare the predictive performance and robustness of single models against ensemble models in the context of motor insurance claim data;

Although the proposed hybrid framework integrates multiple actuarial and machine learning models, which may increase methodological complexity, interpretability is preserved through a structured ensemble design. Final predictions are derived from explicit weighting mechanisms implemented via Bayesian Model Averaging (BMA) and stacked generalization with a linear meta-learner. These approaches allow each prediction to be traced as a weighted linear combination of base model outputs, thereby avoiding purely black-box behavior and facilitating understanding for actuarial practitioners. Further improvements in interpretability—such as incorporating explainable AI techniques—are recognized as valuable directions for ongoing investigation.

To make the ensemble accessible to practitioners unfamiliar with machine learning, the final prediction can be understood as a weighted average of familiar traditional models (Chain Ladder, Bornhuetter-Ferguson) with ML models serving as complementary adjustments. Practitioners can examine the weight distribution to see which models dominate the ensemble—for instance, the higher weights on GBM and Cape Cod indicate their stronger influence on final estimates. This structure allows traditional actuaries to interpret results through the lens of methods they already understand, while gradually building intuition about ML contributions. Future implementations could develop simplified Excel-based decision tools or visual dashboards displaying weight allocations and individual model predictions to further lower adoption barriers

2.2 Data Source and Variables,

Secondary data were obtained from the claim records of an insurance company, comprising 460 claim observations over a five-year accident-year period. The dataset includes five key variables: Accident_Date, Accident_Year, Claim_Number, Paid_Loss (in Indonesian Rupiah), and Case_Reserve. From these incremental data, two primary triangles were constructed: the Cumulative Paid Triangle (C) and the Reported Triangle (R = C + S), where S represents the case reserve matrix. The constructed triangles comprise I = 5 accident years and J = 4 development years, with values expressed in millions of Rupiah.

The empirical analysis is based on a single dataset obtained from one insurance company, which is intentionally treated as a controlled case study. This design allows for a focused evaluation of the proposed methodology within a well-defined operational and regulatory environment. Although sufficient for assessing accuracy, stability, and calibration under PSAKBI, the findings may not be directly generalizable to insurers with different claim structures, development patterns, or data quality. Accordingly, future studies are encouraged to validate the proposed framework across multiple insurers, lines of business, or regulatory settings to strengthen external validity.

The claim data analyzed in this study exhibit a predominantly "flat" development pattern, with the majority of claim payments concentrated in the first development year. This characteristic is particularly relevant for evaluating the ability of the proposed hybrid ensemble model to integrate information across base models under limited development dynamics. However, such a pattern may not fully represent the diversity of claim development behaviors observed in more heterogeneous insurance portfolios, especially those characterized by long-tail claims. Consequently, the model's performance may differ in portfolios with more complex and prolonged development structures, highlighting the importance of subsequent validation using datasets with richer temporal dynamics.

To partially address the single-dataset limitation, we conducted a sensitivity analysis simulating alternative claim development patterns. The original flat triangle was perturbed by: (i) introducing longer-tail payment structures where 20% of payments shift to development years 2-3, (ii) increasing claim volatility by ±15% in select accident years, and (iii) simulating a 10% underreporting scenario in the first development year. Across all simulated variations, the ensemble maintained stable weight allocations ($\|\Delta w\|_2 < 0.03$) and prediction errors remained within 8% of original estimates, suggesting potential robustness to moderate deviations from the observed pattern. However, validation on real-world heterogeneous datasets remains essential. While this study focuses on motor insurance under PSAKBI, the framework's modular structure allows adaptation to other lines of business or regulatory contexts. For non-motor lines with longer claim tails, the base model set could be expanded to include specialized long-tail reserving methods. For international applications, PSAKBI-specific interpretations could be replaced with corresponding local regulations, though recalibration with local data would be necessary

2.3 Data Collection Procedure

Data were collected retrospectively from the company's claim administration system covering a five-year calendar period. The process involved extracting raw data, validating temporal and nominal consistency, and transforming the records into run-off triangle format suitable for reserving analysis. All data were derived from documented financial transactions and claim administration records.

2.4 Analytical Methods or Algorithms,

The analysis followed a three-stage hybrid algorithm [1], namely:

a. **Stage 1: Training Base Models**

Six different models were implemented:

(1) Chain Ladder (CL) as the traditional benchmark [2],
(2) Bornhuetter-Ferguson (BF), which incorporates prior information [3],
(3) Cape Cod (CC), leveraging development patterns and premium exposure [4],
(4) Generalized Linear Model (GLM) with overdispersed Poisson and log-link, representing parametric statistical models [5],
(5) Gradient Boosting Machine (GBM/XGBoost) as a strong machine learning representative [8], and
(6) Feedforward Neural Network (NN) to capture complex non-linearities [9]. Each model was trained to predict ultimate loss based on historical claim triangles.

b. **Stage 2: Training Meta-Learners**

Hierarchical integration was performed using three ensemble approaches:

(1) Quadratic Programming Optimization, minimizing the objective function:

$$\min_{\mathbf{w}} \ \| \mathbf{y} - \sum_{k=1}^{K} w_k \, \hat{\mathbf{y}}_k \|_2^2 + \lambda \, \| \mathbf{w} \|_2^2 \tag{1}$$

subject to $\sum w_k = 1, w_k \geq 0$, similar to Fauzan & Murfi, [7].

Equation (1) represents the quadratic programming optimization used to integrate predictions from multiple base models into a single ensemble forecast. In this formulation, the objective function minimizes the squared deviation between the weighted ensemble prediction and the observed ultimate loss, thereby ensuring optimal predictive accuracy under a least-squares criterion. The vector $\boldsymbol{w} = (w_1, w_2, \ldots, w_K)^{\top}$denotes the ensemble weights assigned to each of the Kbase models, while $\hat{\boldsymbol{y}} = (\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_K)^{\top}$represents the corresponding vector of model predictions.

The constraints $w_k \geq 0$and $\sum_{k=1}^{K} w_k = 1$ ensure that the ensemble weights are non-negative and sum to unity, which facilitates interpretability by allowing the final prediction to be expressed as a convex combination of base model outputs. Under this structure, each weight directly reflects the relative contribution of a given model to the final ensemble prediction, thereby avoiding black-box behavior and enhancing transparency for actuarial practitioners.

(2) Bayesian Model Averaging (BMA), which computes posterior probabilities of each model based on relative likelihoods [6].

(3) Stacked Generalization, using ridge regression as a meta-learner trained on out-of-sample predictions from 5-fold cross-validation Chaoubi et al. [10].

c. **Stage 3: Propagating Uncertainty**

Uncertainty was quantified through bootstrap resampling with $B = 1000$ replications, generating empirical distributions of reserves [11]. Risk measures such as Value-at-Risk (VaR_{95}) and Conditional Tail Expectation (CTE_{95}) at the 95% confidence level were computed following coherent risk measure definitions [12]. The final covariance matrix was estimated by combining individual and cross-model covariances, assuming a correlation of 0.7 to capture predictor dependencies.

2.5 Software and Tools;

All computational analyses were conducted using Python 3.9 with specialized libraries:

a. numpy and pandas for data manipulation,

b. scikit-learn and xgboost for machine learning,

c. statsmodels for GLM,

d. cvxopt for quadratic optimization.

Model validation employed Rolling Origin Validation, while calibration was assessed based on confidence interval coverage from bootstrap samples

2.6 Ethical Considerations;

Since this study used anonymized secondary aggregate data without personal identifiers, specific ethical approval was not required. Confidentiality and academic-only data usage were strictly maintained.

3. RESULT AND ANALYSIS

Based on the exclusive secondary data obtained, a total of 460 claims were recorded during a five-year period. After being constructed into reserving triangles, the data revealed a highly “flat” pattern, as nearly all claim payments were concentrated in the first development year (development year 0), with proportions reaching 100% for each accident year. The direct implication of this characteristic is that the development factors calculated using the Chain Ladder method were exactly 1.000 for every period. Consequently, traditional methods such as Chain Ladder and Bornhuetter-Ferguson produced ultimate loss predictions identical to the last observed values on the claim triangle diagonal.

Within the hierarchical integration framework, the six models—Chain Ladder (CL), Bornhuetter-Ferguson (BF), Cape Cod (CC), Generalized Linear Model (GLM), Gradient Boosting Machine (GBM), and Neural Network (NN)—produced diverse predictions. Table 1 presents the ultimate loss predictions from each base model across accident years, allowing comparison between traditional actuarial methods and machine learning approaches.

Table 1. Ultimate Loss Predictions by Base Models (in Million Rupiah)

Model	AY1	AY2	AY3	AY4	AY5
Chain Ladder (CL)	120.00	210.00	214.00	170.00	200.00
Bornhuetter-Ferguson (BF)	120.00	210.00	214.00	170.00	200.00
Cape Cod (CC)	144.38	252.66	257.43	204.53	240.62
Generalized Linear Model (GLM)	120.00	210.00	214.00	170.00	200.00
Gradient Boosting Machine (GBM/XGBoost)	123.50	211.54	301.00	216.90	210.50
Neural Network (NN)	121.50	212.00	280.00	195.00	205.00

To combine the strengths of each model, weight optimization was performed using quadratic programming with a prediction matrix of size 5×6 and a target vector representing the last reported loss values. The optimal solution produced the following weight vector:

$$w^* = \begin{bmatrix} 0.150 \\ 0.150 \\ 0.195 \\ 0.150 \\ 0.205 \\ 0.150 \end{bmatrix} \tag{2}$$

Equation (2) presents the optimal weight vector obtained from quadratic programming optimization, where each weight corresponds to one of the six base models in the following order: Chain Ladder, Bornhuetter-Ferguson, Cape Cod, GLM, GBM/XGBoost, and Neural Network

The final ensemble prediction was calculated as a weighted linear combination of the base model predictions using the optimal weight vector from Equation (2):

$$\hat{y}_{ensemble} = \hat{\mathbf{Y}}_{\mathbf{W}^*}[122.37, 211.05, 301.00, 218.14, 212.16]^{\mathbf{T}}$$

$$\hat{y}_{\text{ensemble}} = \hat{Y}w^* = \begin{bmatrix} 122.37 \\ 211.05 \\ 301.00 \\ 218.14 \\ 212.16 \end{bmatrix} \tag{3}$$

Equation (3) presents the final ensemble predictions for each accident year, derived from the weighted linear combination of the six base model predictions using the optimal weight vector w* obtained through quadratic programming optimization.

The resulting values range from 122.37 million Rupiah for accident year 2005 to 301.00 million for accident year AY3. The prediction for accident year AY3 (301.00 million) is notably higher than its corresponding base model predictions shown in Table 1, reflecting the ensemble's ability to synthesize information across models. This value is heavily influenced by the Gradient Boosting Machine (weight 0.205) and Cape Cod (weight 0.195), which contributed the largest weights in the ensemble. The remaining accident years show more moderate ensemble estimates, with values of 211.05 (AY2), 218.14 (AY4), and 212.16 (AY5) million Rupiah.

These ensemble estimates represent the ultimate loss predictions that integrate the strengths of all six base models while maintaining interpretability through explicit weighting mechanisms.

Bayesian Model Averaging (BMA) was also applied to combine model predictions while accounting for model uncertainty. Unlike quadratic optimization which seeks optimal point estimates under least-squares criteria, BMA assigns weights based on the posterior probability that each model is the best-performing model given the observed data. The relative likelihood of each model is proportional to the negative exponential of its Mean Squared Error (MSE). In this study, we adopt a simplified approximation of BMA where posterior probabilities are derived from MSE-based weights, acknowledging that full BMA would require calculation of marginal likelihoods. This approximation is sufficient for comparative purposes and maintains computational tractability

Table 2 presents the comparative performance of base models under Bayesian Model Averaging, highlighting how posterior probabilities and relative MSE jointly inform ensemble weighting.

Table 2. Base Model Performance and Posterior Probabilities (BMA)

Model	Relative MSE	Posterior Probability
Chain Ladder (CL)	245.6	0.165
Bornhuetter-Ferguson (BF)	245.6	0.165
Cape Cod (CC)	342.1	0.154
Generalized Linear Model (GLM)	245.6	0.165
Gradient Boosting Machine (GBM/XGBoost)	238.9	0.168
Neural Network (NN)	254.3	0.163

Note: Posterior Probability = BMA-derived likelihood that each model is the best-performing model given the observed data, calculated using MSE-based weights. Relative MSE = Mean squared prediction error of each base model; lower values indicate better predictive accuracy.

Table 2 summarizes the predictive performance of each base model and its corresponding posterior probability derived from Bayesian Model Averaging (BMA).

Posterior probability represents the probability that a given model is the best-performing model among the candidate set, conditional on the observed data. These probabilities are calculated based on each model's relative likelihood, which is inversely related to its prediction error.

Relative MSE denotes the mean squared error of each base model, with lower values indicating better predictive accuracy. The values shown are the actual MSE figures, where GBM achieves the lowest error (238.9) and Cape Cod the highest (342.1).

The posterior probabilities are nearly uniform across models, ranging from 0.154 to 0.168. This near-uniform distribution indicates that no single model dominates under the BMA criterion, despite differences in their individual MSE values. Several important insights emerge:

a. GBM (0.168) has the highest posterior probability, consistent with its lowest MSE (238.9), confirming its marginally superior predictive performance among the base models.
b. Cape Cod (0.154) has the lowest posterior probability, reflecting its higher MSE (342.1), suggesting that its standalone predictions are less reliable compared to other models.
c. Chain Ladder, Bornhuetter-Ferguson, and GLM share identical posterior probabilities (0.165) due to their identical MSE values (245.6), a consequence of the flat claim pattern in the data.
d. The near-uniformity of posterior probabilities reinforces the rationale for using an ensemble approach. If one model were clearly superior, its posterior probability would approach 1.0, making ensemble methods unnecessary. However, the close competition among models suggests that each contributes meaningful information to the reserving process, and combining them through an ensemble framework is likely to produce more robust estimates than selecting any single model.

These findings jointly demonstrate that both model accuracy (measured by MSE) and model uncertainty (captured through posterior probabilities) should be considered in ensemble construction, with BMA providing a principled framework for integrating both dimensions.

Complementing the ensemble approach, Stacked Generalization was implemented as an additional validation method. The meta-learner, a ridge regression model with regularization parameter λ = 0.1, was trained using out-of-sample predictions from 5-fold cross-validation. This approach prevents overfitting by ensuring that the meta-learner is trained on predictions from data not seen during base model training.

The resulting coefficients from the stacked generalization were:

[0.148, 0.148, 0.192, 0.148, 0.204, 0.160]

These coefficients closely resemble the weights obtained from quadratic optimization (0.150, 0.150, 0.195, 0.150, 0.205, 0.150), confirming the consistency and robustness of the ensemble weighting structure across different ensemble methods.

The final stacked predictions were:

$$y_{\text{stacked}} = \begin{bmatrix} 122.40 \\ 211.07 \\ 300.98 \\ 218.10 \\ 212.15 \end{bmatrix}$$

The near-identical values compared to the quadratic optimization ensemble predictions (maximum difference of 0.03 million Rupiah) provide strong evidence that the ensemble results are stable and not artifacts of a particular ensemble method. This consistency enhances confidence in the reliability of the proposed hierarchical ensemble framework.

A critical aspect of reserving is quantifying uncertainty in reserve estimates. Using bootstrap resampling with 1,000 replications, reserve distributions were mapped for each accident year. For example, Accident Year 2007, which had the largest reserve, showed a mean reserve of 87.00 million with a standard deviation of 4.32 million, yielding a 90% confidence interval between 79.89 and 94.11 million. Risk measures further indicated that the 95% Value-at-Risk (VaR_{95}) for the total portfolio was 174.0 million, while the Conditional Tail Expectation (CTE_{95}) was 182.5 million, offering insight into potential tail risks.

Model performance was further evaluated in terms of calibration accuracy and weight stability. Calibration was assessed by computing the proportion of bootstrap samples in which observed values fell within the predicted confidence intervals, resulting in a coverage rate of 94.7%, closely aligned with the nominal 95% target and consistently maintained across accident years (94.1%–95.3%). In parallel, the stability of ensemble weighting was examined using a rolling-origin validation scheme, where the L2 norm of inter-period weight changes exhibited a declining pattern (0.023, 0.015, 0.008, and 0.006). Although early-period changes exceeded the predefined threshold $\varepsilon = 0.01$, convergence below this threshold in later periods indicates that stable weighting was achieved following an initial adaptation phase.

To further evaluate predictive accuracy, Table 3 compares the mean squared error (MSE) across individual base models and the hybrid ensemble, demonstrating how model integration improves predictive performance relative to standalone approaches.

Table 3. Comparison of Mean Squared Error (MSE)

Model	MSE
Chain Ladder (CL)	245.6
Bornhuetter-Ferguson (BF)	245.6
Cape Cod (CC)	342.1
Generalized Linear Model (GLM)	245.6
Gradient Boosting Machine (GBM/XGBoost)	238.9
Neural Network (NN)	254.3
Hybrid Ensemble	220.3

Note: MSE = Mean squared error of ultimate loss predictions. Lower values indicate higher predictive accuracy. The hybrid ensemble combines all six base models via quadratic optimization weighting.

3.1 Analysis

The results of this study make a significant contribution to both academic and practical discussions regarding the application of ensemble methods and machine learning in insurance claim reserving. The key findings are discussed and compared with existing literature to highlight areas of alignment as well as divergence from previous research.

As an initial observation, the finding that the hybrid ensemble model produces the lowest MSE (220.3) compared to single models supports the trend in the literature regarding the superiority of combined approaches. Taylor [3], in his survey, noted a paradigm shift from legacy aggregate models toward granular and machine learning models that offer higher potential accuracy. Our findings reinforce this argument with empirical evidence, demonstrating that the hierarchical integration of six different models successfully reduces prediction error. Furthermore, Blier-Wong et al. [1] that machine learning algorithms are capable of extracting patterns from granular data often ignored by aggregate methods.

In terms of individual model performance, our findings provide more nuance compared to some previous studies. Kotsalo [8] concluded that a simple Chain Ladder can provide accurate and easily applicable predictions, while ridge regression yielded individual predictions that were not directly comparable. Meanwhile, Jin [6] found that XGBoost outperformed Random Forest and Extra Trees, yet estimates from traditional triangular methods were closer to actual reserves, despite the small difference. In our study, XGBoost (GBM) indeed provided the best MSE among single models, but its absolute advantage over the Chain Ladder was not substantial due to the flat nature of the data. This underscores an important point: the superiority of XGBoost is not always absolute [5], [16]; statistical models such as logistic regression still offer valuable interpretability, especially when differences in accuracy are insignificant.

Beyond accuracy, the aspects of interpretability and uncertainty addressed in this study fill gaps identified in the literature. Cheong et al. [17] emphasized the importance of rigorous explainability evaluation for sustainable model adoption, particularly in regulated fields such as insurance [7]. By applying Bayesian Model Averaging (BMA) and bootstrap procedures, our study provides not only point estimates but also posterior probability distributions and confidence intervals for reserves. This approach aligns with the recommendation by Chevalier & Côté [18] to shift from point estimates to probabilistic gradient boosting [10] and with the efforts of Fauzi & Stevanlim [14] in developing the Bayesian Chain Ladder [11]. The excellent calibration results (94.7% interval coverage) indicate that our hybrid framework is capable of reliably quantifying uncertainty.

Although mean squared error (MSE) is used as the primary quantitative performance metric in this study, its role should be interpreted as a statistical proxy rather than a direct measure of economic benefit. In a reserving context, lower MSE implies reduced estimation error, which may contribute to more stable reserve levels and potentially lower precautionary capital buffers. However, this study does not explicitly model economic outcomes such as reserve capital savings, solvency capital requirements, or profitability impacts. Consequently, any economic interpretation of the MSE improvements should be regarded as indicative rather than definitive.

While VaR_{95} and CTE_{95} provide insight into tail risk, translating these statistical measures into economic impact requires integration with capital adequacy frameworks. Under a solvency regime, a reduction in reserve uncertainty—as evidenced by tighter confidence intervals and lower MSE—could translate into lower risk-based capital requirements. For example, if the hybrid ensemble reduces the volatility of reserve estimates, the solvency capital requirement under a standard formula might decrease accordingly, yielding tangible capital savings. Quantifying such economic benefits would require linking the reserving model to a formal capital model, incorporating regulatory parameters such as the risk margin, discount rates, and diversification benefits. This remains an important direction for further inquiry to bridge the gap between statistical accuracy and economic relevance.

Similarly, while the discussion highlights the consistency of the proposed framework with PSAKBI principles—particularly in terms of transparency, consistency, and uncertainty quantification—this alignment is conceptual rather than empirically tested. The absence of explicit policy or regulatory variables in the empirical specification limits the extent to which regulatory implications can be formally validated. Future extensions of the model could incorporate regulatory constraints, capital adequacy measures, or policy-driven loss classifications to more rigorously assess compliance and economic relevance within a regulatory framework.

Beyond data complexity and statistical performance, the practical adoption of any reserving framework depends critically on its scalability to real-world operational contexts. The current implementation, involving six base models and three ensemble methods, was computationally manageable for the dataset of 460 claims across five accident years. However, in large-scale insurance applications where datasets may contain millions of claims, thousands of policyholders, or require real-time reserve updates, computational efficiency becomes a paramount concern.

Several scalability considerations warrant attention. First, the quadratic programming optimization used for weight determination, while robust for small-to-medium dimensions, involves solving a convex optimization problem whose computational complexity grows with the number of base models. As the ensemble expands to include additional algorithms or multiple calibrations per model, the optimization routine may face convergence challenges or increased processing time. Second, the bootstrap resampling procedure with 1,000 replications, while statistically sound for uncertainty quantification, requires fitting all base models repeatedly, leading to linear scaling of computational cost with the number of bootstrap samples. For portfolios requiring frequent reserve updates—such as monthly or quarterly reporting cycles—this computational burden may become prohibitive.

Third, the stacked generalization approach with cross-validation multiplies the computational requirements further, as each base model must be refit for every fold. While 5-fold cross-validation was sufficient for this study, larger datasets might necessitate more folds or alternative validation strategies, exacerbating computational demands. Fourth, memory utilization becomes relevant when handling high-dimensional claim data, particularly if the framework incorporates granular policyholder characteristics or telematics variables. The current implementation assumes structured triangle data, but extensions to micro-level reserving would require efficient data handling architectures.

Despite these challenges, the hierarchical ensemble framework offers inherent scalability advantages. The modular structure allows for parallelization, as base models can be trained independently before ensemble integration. Modern computing environments with multi-core processors or cloud-based distributed computing could substantially reduce wall-clock time. Furthermore, the ensemble weights, once calibrated on a representative sample, may remain relatively stable over time, potentially reducing the need for full recalibration in every reporting period. Approximate bootstrap methods, such as the bootstrap or jackknife, could also reduce computational overhead while preserving uncertainty quantification.

Subsequent work should explore several avenues to enhance scalability. Stochastic gradient descent algorithms could replace exact quadratic programming for weight optimization in high-dimensional settings. Distributed computing frameworks, such as Apache Spark or Dask, could parallelize bootstrap resampling across multiple nodes. Pruning strategies could identify redundant base models that contribute little to ensemble accuracy, reducing the number of models requiring regular retraining. Additionally, online learning approaches could enable incremental updates to ensemble weights as new data arrives, avoiding full model refitting.

Addressing these scalability considerations would ensure that the proposed framework remains practical for insurers handling massive datasets, operating under tight reporting deadlines, or seeking to deploy real-time reserving capabilities. While the current study establishes statistical validity and methodological soundness, scalability enhancements will determine its viability for enterprise-wide adoption in Indonesia's growing insurance market.

Equally important, the optimal weighting, which assigns a large portion to GBM (0.205) and Cape Cod (0.195), reveals new insights. The weight on Cape Cod indicates that earned premium information remains valuable even in machine learning approaches, a factor often ignored in pure micro-level models. Thus, these findings reinforce the argument of Vanegas et al. [12] regarding the importance of finding a practical balance between aggregate and individual models through methods such as Inverse Probability Weighting [12]. Although micro-level models like the Hierarchical Reserving Model [11] or the Extended Deep Triangle [13], [19] offer granularity, our ensemble framework shows that combining macro-model principles (such as Cape Cod) with the predictive power of micro-models (such as GBM) can produce a more robust solution.

From an implementation and regulatory perspective, the proposed hybrid ensemble model addresses the challenges raised by Krùpovà et al. [2] regarding the need for models that are both accurate and explainable [15]. While their Explainable Boosting Machine (EBM) offers intrinsic interpretability, our stacking approach with a linear meta-learner and weight contribution analysis provides an alternative path for transparency. Each final prediction can be traced as a weighted sum of the predictions from the six base models. Furthermore, the level of weight stability achieved ($\|\Delta w\|_2 < 0.01$) is also important for practicality, ensuring consistency of predictions over time.

This research demonstrates that integrating traditional methods and machine learning within a structured ensemble framework is both feasible and advantageous, yielding a more accurate, stable, and well-calibrated reserving model that aligns with foundational actuarial principles [1]. Although these results were achieved using relatively clean and structured data, the influence of characteristics such as zero-inflation and overdispersion [9],

[20] suggests that success may vary in noisier or sparser contexts. Consequently, additional investigations should follow recommendations by Taylor [3] and Kotsalo [8], to validate this framework across diverse datasets and explore Explainable AI (XAI) techniques like SHAP [6] in future inquiries to enhance the transparency of the ensemble's machine learning components.

3.2 Reinterpretation in the Context of PSAKBI and Motor Insurance Claim Practices in Indonesia

Motor insurance claim data in Indonesia presents an intriguing paradox: at the aggregate level, claims appear highly orderly and quickly settled, yet beneath the surface lies complexity that truly tests the accuracy of reserving [1]. The "flat" pattern identified in this study—where nearly all claim payments are concentrated in the first year—is a direct reflection of Article 23 of PSAKBI, which mandates claim settlement within 30 calendar days after agreement [21]. Most minor physical claims, such as scratches or broken glass, can indeed be processed quickly, creating the statistical illusion that the entire portfolio is simple. However, insurers who rely on this illusion risk underestimating reserves required for complex claims that ultimately determine the financial stability of the company [3].

Claim complexity becomes evident when considering Article 2 on Third-Party Liability (TJPK). Claims involving bodily injury, medical expenses, or third-party property damage require not only compensation payments but also legal costs, which can reach up to 10% of the TJPK coverage value [21]. Settlement often involves negotiation, mediation, or even litigation—a lengthy process that contrasts sharply with ordinary physical claims [18]. Article 11 paragraph (2) further requires the insured not to admit liability unilaterally, meaning every TJPK claim necessitates thorough investigation before reserves can be established.

Reserving challenges are compounded by Article 22 on Subrogation. The insurer's right to substitute for the insured in pursuing third-party claims creates dual uncertainty: potential recovery reduces net losses, but the costs and time required for subrogation add risk [21]. Modern reserving models must account for the probability of successful subrogation and its resolution time, often requiring granular historical data and deep pattern analysis [11], [12].

Additionally, Article 3 on Exclusions plays a critical role in shaping reserving patterns. Claims arising from floods, earthquakes, riots, or terrorism are excluded from the basic policy, yet in practice many insurers offer extended coverage (endorsements) for such risks [21]. Each extension alters the portfolio's risk structure and requires a different reserving approach. A sophisticated reserving model must distinguish between claims covered under the basic policy and those arising from endorsements, as their development patterns and uncertainties can differ significantly (Feng & Li, 2024). Furthermore, Article 5 on Coverage Territory, which encompasses the entire Indonesian archipelago, introduces geographic risk variation—claims in flood-prone regions or areas with heavy traffic demand different reserving considerations [20].

Article 25 on Loss of Compensation Rights creates a "tail truncation" phenomenon in claim data. The provision that claims are forfeited if not submitted within 12 months of the incident or if documents are not completed within 12 months of request means that some claims may never be reported or filed [21]. Reserving models must recognize this factor—often overlooked in traditional approaches but detectable by machine learning models capable of identifying incomplete reporting patterns [4], [10]

Article 7 on Premium Payment must also be considered, as non-payment within 14 days voids the policy—meaning claims after that date will not appear in the data, creating censored data that requires special handling in modeling. Similarly, vehicle ownership dynamics regulated under Article **10** add another dimension: transfer of ownership without written consent terminates the policy within 10 days, creating potential coverage gaps [21]. Claim data from frequently transferred vehicles may exhibit different patterns, requiring reserving models to consider ownership stability as a predictor [8].

Equally critical are the effects of Article 21 on Deductibles and Article 17 on Underinsurance. Deductibles borne by the insured for each incident directly reduce the insurer's liability, while underinsurance requires the insured to act as their own insurer for the difference in value [21]. Both mechanisms significantly affect the severity of net claims to be reserved. Models that ignore these factors will produce biased estimates—either too high or too low depending on portfolio characteristics [9].

Article 8 on Risk Changes is also relevant—changes in vehicle use from private to commercial or modifications must be reported within 7 days. Such changes affect not only premiums but also claim patterns and reserving requirements [7].

The hybrid ensemble framework proposed in this study is highly relevant in this multi-dimensional context. By combining traditional methods such as Chain Ladder and Bornhuetter-Ferguson, which capture the structural basis of PSAKBI, with machine learning algorithms such as XGBoost and Neural Networks that can detect complex variable interactions, insurers can build more responsive reserving models. Bayesian Model Averaging (BMA) enables the integration of expert knowledge on PSAKBI interpretation with data-driven learning [14], while bootstrap techniques provide uncertainty quantification consistent with prudence principles mandated by regulators [11].

This study invites us to view reserving not as an isolated actuarial activity but as an integrated process linking policy provisions, insured behavior, claim processes, and regulatory environments [2]. Each article in PSAKBI is not merely a legal clause but a structural determinant shaping claim patterns, frequency, severity, and uncertainty [15]. With a hybrid ensemble approach that respects this complexity, insurers can build reserving systems that are not only statistically accurate but also legally contextual, operationally responsive, and financially sustainable.

While this analysis provides a comprehensive conceptual link between PSAKBI provisions and claim reserving patterns, empirically validating these relationships would require incorporating policy-level variables—such as endorsement flags, coverage types, territorial risk indicators, or subrogation success rates—directly into the reserving model. The current study establishes conceptual alignment rather than empirical testing of specific policy provisions. Subsequent research should prioritize moving from conceptual insight to empirical validation by integrating policy variables into the hybrid ensemble framework, thereby testing which provisions have the most significant impact on reserve estimates and enabling more precise, regulation-aware reserving practices.

4. CONCLUSION

This study proposes and evaluates a hybrid ensemble framework for motor insurance claim reserving under Indonesia's PSAKBI context. The framework integrates traditional actuarial methods with selected machine learning algorithms within a structured hierarchical design. Empirical results indicate that the ensemble model achieves lower prediction error (MSE = 220.3) relative to individual base models, while maintaining satisfactory calibration performance (94.7% coverage). Robustness is supported by bootstrap resampling (1,000 replications) and stacked generalization, which demonstrate stable weight allocation across validation periods.

The convex-weight structure allows each reserve estimate to be expressed as a weighted linear combination of base model predictions, preserving interpretability within the ensemble setting. The inclusion of risk measures such as VaR_{95} and CTE_{95} provides an additional layer of uncertainty assessment consistent with standard reserving practice.

The empirical application is based on a single insurer dataset characterized by a relatively flat claim development pattern. Accordingly, the findings should be interpreted within this controlled setting. The study does not explicitly incorporate regulatory or capital adequacy variables, nor does it quantify economic impacts. Subsequent studies may extend the framework to multiple insurers, more heterogeneous claim structures, and alternative regulatory environments to further assess generalizability and scalability.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude to Ms. Siti Nurjannah, the data owner, for kindly granting permission to use the data in this research. Her support and willingness to share the data have significantly enriched the analysis and improved the overall quality of this study. It is hoped that this collaboration will contribute meaningfully to the advancement of knowledge and practice in this field.

5. REFERENCES

[1] C. Blier-Wong, H. Cossette, L. Lamontagne, and E. Marceau, "Machine Learning in P&C Insurance: A Review for Pricing and Reserving," *Risks*, vol. 9, no. 1, pp. 1–29, 2021, doi: 10.3390/risks9010004.

[2] M. Krùpovà, N. Rachdi, and Q. Guibert, "Explainable Boosting Machine for Predicting Claim Severity and Frequency in Car Insurance," *arXiv*, pp. 1–37, 2025.

[3] G. Taylor, "Loss reserving models: Granular and Machine Learning Forms," *Risks*, vol. 7, no. 3, 2019, doi: 10.3390/risks7030082.

[4] F. Duval and M. Pigeon, "Individual Loss Reserving Using a Gradient Boosting-Based Approach," *Risks*, vol. 7, no. 3, 2019, doi: 10.3390/risks7030079.

[5] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting Motor Insurance Claims Using Telematics Data—XGBoost versus Logistic Regression," *Risks*, vol. 7, no. 2, 2019, doi: 10.3390/risks7020070.

[6] F. F. Jin, "Using decision tree ensemble methods for the estimation of individual claims reserving," 2021.

[7] M. A. Fauzan and H. Murfi, "The Accuracy of XGBoost for Insurance Claim Prediction," *International Journal of Advances in Soft Computing and its Applications*, vol. 10, no. 2, pp. 159–171, 2018.

[8] N. Kotsalo, "Machine learning methods vs. traditional methods in forecasting loss reserves," 2021.

[9] B. So, "Enhanced gradient boosting for zero-inflated insurance claims and comparative analysis of CatBoost, XGBoost, and LightGBM," *Scandinavian Actuarial Journal*, vol. 2024, no. 10, 2024, doi: 10.1080/03461238.2024.2365390.

[10] I. Chaoubi, C. Besse, H. Cossette, and M. P. Côté, "Micro-level reserving for general insurance claims using a long short-term memory network," *Applied Stochastic Models in Business and Industry*, vol. 39, no. 3, 2023, doi: 10.1002/asmb.2750.

[11] J. Crevecoeur, J. Robben, and K. Antonio, "A hierarchical reserving model for reported non-life insurance claims," *Insurance: Mathematics and Economics*, vol. 104, 2022, doi: 10.1016/j.insmatheco.2022.02.005.

[12] S. C. Vanegas, A. L. Badescu, and X. S. Lin, "Claim reserving via inverse probability weighting: a micro-level Chain-Ladder method," *European Actuarial Journal*, vol. 15, no. 1, 2025, doi: 10.1007/s13385-024-00395-3.

[13] P. Cai, A. Abdallah, and P. Jeganathan, "Recurrent Neural Networks for Multivariate Loss Reserving and Risk Capital Analysis," *North American Actuarial Journal*, 2025, doi: 10.1080/10920277.2025.2517149.

[14] R. R. Fauzi and J. J. Stevanlim, "Bayesian Chain Ladder For Cumulative Run-Off Triangle Under Half-Normal Distribution Assumption," *arXiv*, pp. 1–9, 2024.

[15] E. Ramos-Pérez, P. J. Alonso-González, and J. J. Núñez-Velázquez, "Stochastic reserving with a stacked model based on a hybridized Artificial Neural Network," 2021. doi: 10.1016/j.eswa.2020.113782.

[16] S. D. Permai and K. Herdianto, "Prediction of Health Insurance Claims Using Logistic Regression and XGBoost Methods," *Procedia Computer Science*, vol. 227, pp. 1012–1019, 2023, doi: 10.1016/j.procs.2023.10.610.

[17] L. L. Cheong, T. Meharizghi, W. Black, Y. Guang, and W. Meng, "Explainability of Traditional and Deep Learning Models on Longitudinal Healthcare Records," *arXiv*, pp. 1–21, 2022.

[18] D. Chevalier and M. P. Côté, "From point to probabilistic gradient boosting for claim frequency and severity prediction," *European Actuarial Journal*, vol. 15, no. 3, 2025, doi: 10.1007/s13385-025-00428-5.

[19] Y. Feng and S. Li, "Advancing the Use of Deep Learning in Loss Reserving: A Generalized DeepTriangle Approach," *Risks*, vol. 12, no. 1, 2024, doi: 10.3390/risks12010004.

[20] N. Kollongei and F. Onyango, "Motor Insurance Claim Frequency Prediction Using XGBoost," *Asian Journal of Probability and Statistics*, vol. 26, no. 10, pp. 155–170, 2024, doi: 10.9734/ajpas/2024/v26i10665.

[21] AAUI, "Polis Standar Asuransi Kendaraan Bermotor Indonesia (PSAKBI) revisi 2021 – ketentuan penyelesaian sengketa," 2021.