IJIRSS
International Journal of Innovative Research & Scientific Studies

ISSN: 2617-6548

URL: www.ijirss.com

Hyperparameter optimisation of generalised linear models in vehicle insurance pricing

Sandile Buthelezi[1*], Taurai Hungwe[2], Solly Matshonisa Seeletse[1], Vimbai Mbirimi-Hungwe[3]

[1]Department of Computer Science and Information Technology, Pretoria, South Africa.
[2]Department of Statistical Sciences, Pretoria, South Africa.
[3]Department of Academic Literacy and Science Communication Sefako Makgatho University of Health Sciences, Pretoria, South Africa.

Corresponding author: Sandile Buthelezi (*Email: jsbuthelezi@gmail.com*)

Abstract

Predictive modelling is increasingly important in data-driven decision-making, yet traditional statistical approaches such as the Generalised Linear Model (GLM) with a Gamma distribution often exhibit limited accuracy when applied to complex datasets. This study compared the performance of the standard GLM with an optimised iterative hybrid approach employing machine learning algorithms, including XGBoost, RF, Gradient Boosting GBM), and Artificial Neural Networks (ANN). Models were trained and tested on the same dataset, and performance was assessed using three metrics: coefficient of determination (R^2), root mean squared error (RMSE) and mean absolute error (MAE). Hypothesis testing was conducted using t-tests and F-tests at a 5% significance level. Results showed that the GLM baseline achieved modest explanatory power (training $R^2 = 0.23$; test $R^2 = 0.19$) and comparatively high prediction errors (RMSE ≈ 83,992–84,789; MAE ≈ 56,943–60,897). In contrast, hybrid machine learning models performed substantially better, with XGBoost, RF, and ANN each achieving $R^2 = 0.42$ on the test set, RMSE values around 81,200, and competitive MAE scores. Statistical testing confirmed significant improvements in R^2 and RMSE, while MAE differences were less conclusive under unequal variance assumptions. These findings highlight the limitations of conventional GLMs and the enhanced generalisability of hybrid methods. In conclusion, the optimised iterative hybrid approach offers a more reliable and accurate predictive framework. It is recommended that organisations adopt hybrid models, particularly XGBoost and ANN, for predictive tasks requiring high levels of accuracy, while future research should investigate issues of interpretability, computational efficiency, and scalability in applied context.

Keywords: GLMs, Hyperparameter, Machine Learning, Optimisation, Pricing.

DOI: 10.53894/ijirss.v9i1.11199
Funding: This study received no specific financial support.
History: Received: 27 November 2025 / **Revised:** 2 January 2026 / **Accepted:** 6 January 2026 / **Published:** 23 January 2026

Competing Interests: The authors declare that they have no competing interests.
Authors' Contributions: All authors contributed equally to the conception and design of the study. All authors have read and agreed to the published version of the manuscript.
Transparency: The authors confirm that the manuscript is an honest, accurate, and transparent account of the study; that no vital features of the study have been omitted; and that any discrepancies from the study as planned have been explained. This study followed all ethical practices during writing.
Publisher: Innovative Research Publishing

1. Introduction

Accurate risk assessment is a fundamental aspect of vehicle insurance pricing, enabling insurers to determine premiums that accurately reflect the underlying risk profiles of policyholders. GLMs have traditionally been the standard methodology in insurance due to their interpretability and statistical robustness. However, they often struggle to capture complex, nonlinear relationships within insurance data, resulting in suboptimal predictive performance Greberg and Rylander [1]. Xie and Shi [2] highlight that while GLMs facilitate feature selection more effectively than decision-tree-based techniques or neural networks (NNs), they lack the capacity to identify intricate or nonlinear interactions between risk factors and the response variable, thereby affecting pricing accuracy. Consequently, this limitation necessitates the exploration of alternative approaches, including advanced nonlinear models.

Similarly, Havrylenko and Heger [3] emphasise that the effectiveness of GLMs in insurance companies is contingent on variable selection, expert judgement, and visual performance indicators, rendering the model selection process time-consuming and resource intensive. With the advancement of ML, new opportunities have emerged to enhance traditional actuarial models by integrating data-driven techniques that improve prediction accuracy and model robustness. For instance, Buthelezi, et al. [4] investigated the superiority of twenty-two models for pricing automobile insurance, ranging from traditional actuarial methods to modern statistical models such as ML algorithms. Their study explores the evolving landscape of risk factors and market dynamics, highlighting the potential benefits of leveraging these advanced methods. The findings indicate that ANNs, NNs, XGBoost, and RF outperform traditional models, demonstrating that modern statistical techniques can estimate risk exposure more accurately than conventional GLMs.

However, most of these models are inherently complex and lack transparency in their decision-making processes, leading to their classification as 'black box' models [5]. Despite these advancements, GLMs remain preferred due to their ease of interpretability and ability to provide a clear understanding of how each predictor influences the outcome for pricing [6]. Addressing the challenge of identifying interactions between variables particularly in datasets with many predictors, where manual selection is time-consuming and heavily reliant on actuarial expertise requires an automated approach. Havrylenko and Heger [3] propose the use of hybrid NN-GLM, a model-specific interaction detection method to optimise claim count predictions efficiently. Wilson, et al. [6] compares GLMs with GBM and NN for predicting loss costs in motor insurance. The findings indicated that NN models, particularly a hybrid model combining GLM predictions with NN, outperform traditional GLMs and GBMs in predictive accuracy

In line with these developments, this study explores the hybrid optimisation of GLMs through a comparative analysis of model performance, evaluating XGBoost, RF, GBM, and ANN for claim amount predictions. This research assesses the effectiveness of hybrid GLM-ML approaches against traditional GLMs. By bridging the gap between conventional statistical modelling and modern ML techniques, this study highlights the potential of hybrid approaches to transform vehicle insurance pricing. The results underscore the value of integrating ML with actuarial methods, paving the way for more precise, data-driven insurance models that enhance fairness, efficiency, and competitiveness in the industry.

2. Literature Review

The advancement of ML algorithms continues to demonstrate superiority in enhancing the predictive accuracy of pricing in motor insurance. However, Panjee and Amornsawadwatana [7] conducted a study comparing predictive modelling approaches for claim frequency and severity in cross-border cargo insurance. Their research identified the optimal modelling approach between GLMs and advanced ML techniques. The findings revealed that XGBoost is a robust predictor for claim frequency, whereas the GLM (Gamma) model outperforms both XGBoost and GBM in severity modelling.

Similarly, Clemente, et al. [8] examined the predictive performance of the GBM in comparison to the standard GLM within a Poisson claim frequency framework. Their findings indicated that GBM outperformed the classical GLM in predicting claim frequency. However, in modelling claim severity, the traditional GLM demonstrated superior performance over GBM.

Holvoet, et al. [9] further contributed to this body of research by benchmarking the performance of GLMs against GBMs and feed-forward neural networks (FFNNs) in insurance pricing. Their study emphasised that, while ML techniques such as FFNNs can capture complex patterns within the data, GLMs remain valuable due to their interpretability.

Collectively, these studies illustrate the evolving landscape of predictive modelling in insurance. While ML techniques have proven effective in enhancing predictive accuracy, traditional GLMs continue to offer valuable insights, particularly in terms of interpretability and simplicity. Ardabili, et al. [10] further emphasise that ML algorithms are continuously advancing, incorporating novel learning methods that drive rapid evolution. The development of ML models increasingly leverages hybridisation and ensemble techniques, enhancing their computational efficiency, functionality, robustness, and predictive accuracy. Although numerous hybrid and ensemble ML models have been introduced, they have not been systematically surveyed in a comprehensive manner.

Using the archival technique, Table 1 presents several studies where GLM has been hybridised with advanced statistical methods to enhance predictive accuracy, supporting the statement by Ardabili, et al. [10]. The table highlights recent studies from the period of 2016 to 2025 showcasing the evolving role of ML and hybrid modelling approaches in insurance pricing, particularly in predicting claim frequency and severity. However, studies specifically focusing on claims prediction, particularly in terms of claim losses, remain limited, thus underscoring the need for further research in this area.

Table 1.
The studies that showcase crossbreed models against the GLMs.

Case Study	Type of Publication	Author(s)	Article Contribution
Claims frequency and Severity	Journal paper	Wilson, et al. [6]	The study explores insurance pricing in the motor insurance industry using ML methodologies like GLM, GBM, and ANN, revealing a hybrid model with better predictions, emphasising the need for alternative modelling approaches.
Claims frequency	Journal paper	Brauer [11]	The paper presents novel methods to enhance actuarial non-life models using transformer models for tabular data. Building on the foundation of combined actuarial neural networks and localGLMnet, the methods achieve better results than benchmark models while preserving the structure of the underlying actuarial models.
Claims frequency	Journal Paper	Havrylenko and Heger [3]	The quality of GLMs in insurance companies depends on the choice of interacting variables. An automated approach using NN, and a model-specific interaction detection method is proposed to improve predictive power.
Claims frequency and Severity	Journal paper	Van Oirbeek, et al. [12]	The study proposes a novel application of the Genetic Algorithm (GA) to efficiently identify main and interaction effects in GLMs, even in high variable count scenarios. The GA aligns GLM predictions with black ML models, enhancing interpretability and reliability.
Claims frequency and Severity	Journal paper	Holvoet, et al. [9]	The proposed approach involves training a combined actuarial neural network (CANN), quantifying the strength of each pairwise interaction, ranking them using a neural interaction detection (NID) algorithm, and analysing top-ranked interactions using mini-GLMs to identify the next-best interaction to be included in the benchmark GLM.
Claims frequency	Journal paper	Novkaniza, et al. [13]	The study employs the Poisson-Gamma Hierarchical Generalized Linear Model (PGHGLM) to calculate accurate vehicle insurance premium rates, demonstrating its practical application in generating age-specific premiums and enhancing fairness.
Severity	Journal paper	Seyam and Elsalmouny [14]	The Misr Insurance Company in Egypt, the largest insurance company in Egypt, is proposed to estimate the pure premium using alternative tariff systems. The system constructs insurance rates based on risk factors, using three statistical models: GLM, Generalized Linear Mixed Model (GLMM), and Generalized Additive Model (GAM). However, the research found that GLMM is the most convenient model for ratemaking
Claims frequency and Severity	Thesis	Berry [15]	The study presents a new method for estimating automobile insurance premiums using hidden Markov models (HMM). It combines a Poisson-gamma HMM and a hybrid between HMMs and HMM-GLM. The models address overdispersion in claim counts and introduce dependence between severity and claim count. Simulations show HMM-GLM outperforms standard GLM in some cases.
Claims frequency and Severity	Thesis	Reil [16]	The thesis investigates the use of ML in non-life insurance pricing, comparing it to GLMs. Using XAI techniques, it found ML models outperform GLMs in predictive power for both severity and frequency claims data. However, ensuring model explainability remains a challenge. The study proposes a hybrid approach, combining ML and GLMs to improve accuracy without compromising interpretability.

3. Methodology

The methodology of the study includes data collection, hyperparameter tuning, and the application of both machine learning (ML) and traditional statistical methods such as XGBoost, Artificial Neural Networks (ANN), Gradient Boosting Model (GBM), and Random Forest (RF). Additionally, it details the risk evaluation and model assessment using root mean square error (RMSE), coefficient of determination (R^2), and mean absolute error (MAE).

3.1. Data Collection

The study uses historical claims data from a South African non-life vehicle insurance company from 2021-2024, encompassing 23 variables and categorising incidents into disaster groups and subgroups. The data is only publicly available upon request and withheld from the insurer's name for ethical reasons.

Table 2.
The claims data used in the study with 26 variables explained.

Variable	Explanation
Claim Amount	Amount claimed by the insured/ Policyholder
Premium	Premium paid by the insured/Policyholder
Vehicle Make	The type of car model e.g., VW, Toyota
Vehicle Model	The model of the car e.g., VW 1.2 Trendline
Type of the vehicle	The variable that describes the type of the vehicle e.g., convertible, sedan ...etc
Age of the Vehicle	The variable that describes how old is the vehicle 6 years old
Wheels	Two-wheeler or four-wheeler car e.g., 2WD
Engine size	The size of the engine in the car e.g., 1201 - 1400
Fuel type	Diesel or Unleaded e.g., Diesel
Aspiration Adjustment	The vehicle has a Turbo Yes or No e.g., Yes
Manual or Automatic	The variable that describes if the car is manual or automatic drive e.g., Automatic
Power to Weight Ratio (PtWr)	The ratio of power over the mass of the vehicle classified in CAT1 to CAT 6, CAT1 been the smallest ratio and CAT6 been the highest and e.g., CAT4
Night Parking	The place where the car is parked at night e.g., Locked Garage
Day Parking	The place where the car is parked during the day e.g., Yard/open parking - with Locked Gates/access control/electronic access
Province/City	The region in which the car is driven e.g., Limpopo
Driver Age	The age of the driver e.g., 30
Drive Gender	Male / Female
Licence Type	The type of licence the driver is allowed to drive according to the vehicle weight e.g., B
Claim Incident	Type of claim that led to the incident e.g., storms
Disaster Group	Is the Incident Technological or Natural
Disaster Subgroup	The subgroup of the incident classified as Climatological, Geophysical, Hydrological, Malicious damage, Meteorological and Miscellaneous accident e.g., Miscellaneous accident
Start Date	The date on which the policy was in force/activated e.g., 12/06/2021
End Date	Date in which the policy is expected to expire or end e.g., 12/06/2022
Claim Date	Date of the claim incurred 124/09/2021

3.2. Hyperparameter Tuning

The crossbreeding method incorporates hyperparameter tuning, employing the grid search optimisation technique within ML algorithms. Following this process, the selected features were identified that enhance the algorithm’s accuracy. These features were subsequently deployed in a GLM with a Gamma distribution, where accuracy was observed. This step aimed to reduce the time required by pricing specialists to identify the most interactive variables manual that improve the accuracy of the GLM.

Hyperparameter tuning, also referred to as hyperparameter optimisation, involves selecting the most suitable set of hyperparameters for a ML algorithm to enhance its performance on a particular task. Unlike model parameters, which are learned during the training phase, hyperparameters are predefined and dictate the learning process itself. Effective hyperparameter tuning seeks to identify the combination that minimises a predefined loss function, thereby improving the model's ability to generalise to unseen data [17]. Common techniques for hyperparameter optimisation include grid search, random search, and Bayesian optimisation. Grid search methodically evaluates all possible combinations within a specified subset of the hyperparameter space, while random search selects combinations randomly, potentially exploring a wider range of configurations [18]. Bayesian optimisation constructs a probabilistic model of the objective function and utilises it to iteratively select promising hyperparameter settings. The selection of appropriate hyperparameters is critical, as they can significantly influence the model's accuracy and computational efficiency [17].

3.3. Generalised Linear Model

The GLM frameworks have been identified for non-life insurance pricing as the industry standard. Such frameworks are specified as an extension of the framework of the probability distribution linear model, which is derived from the exponential family and was illustrated by Nelder and Wedderburn [19]. These models seek to estimate an interesting variable (Y) from a set of explanatory variables (X). The GLMs are composed of three parts:

- The first assumption is that an outcome variable (Y) belongs to an exponential family of distributions. This distribution family includes the normal, binomial, Poisson, and gamma distributions. Furthermore, it follows that random factors are independent $(Y_1 \ldots \ldots \ldots Y_n)$ have a similar distribution. Hoscedasticity is commonly assumed in the field of regression.; nevertheless, GLMs are designed to handle heteroscedasticity given by:

$$f(y_i|\theta_i,\phi) = exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i,\phi)\right) \qquad y_i \in P \tag{1}$$

Where p represents the subassembly that is a part of $\mathbb{N}$ or $\mathbb{R}$ set, $\theta_i \in \theta$ the natural parameter and ϕ is the scale parameter.

Then the probability density for the variable $Y_1, Y_2, \cdots, Y_n$ is given by the following expression:

$$f(y|\theta,\phi) = \prod_{i=1}^{n} f(y_i|\theta_i,\phi) \tag{2}$$

By applying the necessary mathematical manipulation, the equation 4.10 can be rewritten by:

$$f(y|\theta,\phi) = exp\left(\frac{\sum_{i=1}^{n} y_i\theta_i - \sum_{i=1}^{n} b(\theta_i)}{\phi} + \sum_{i=1}^{n} c(y_i,\theta_i)\right) \tag{3}$$

- A linear predictor, which has the familiar form of an ordinary linear model.

$$\varphi_i = \alpha + \beta_i x_{i_1} + B_2 x_{i^2} + \cdots + B_k x_{ik} \tag{4}$$

- Given the parameters $\beta_1, \beta_2, \cdots, B_k$ through the function (g) of the mean (μ) written in the linear form for a variable **X**

$$g(\varphi_i) = \boldsymbol{X}\beta = \beta_0 + \sum_{j=1}^{n} B_{i_1} x_{ik} + \epsilon \tag{5}$$

The function (g) is known as a link function given that it is a monotonous and differentiable f$unction$ (g) connected to the linear predictor, mean (μ) and error ϵ. The link function transforms the expected value of the response variable $\mu_i = E[y_i|\{x_1, \dots, x_n\}]$ and since it is invertible, we then get:

$$\mu_i = g^{-1}(\varphi_i) \tag{6}$$

The GLM can be a linear or nonlinear regression model that transforms the expected outcome of a response variable. Furthermore, the conditional variance of a distribution in the exponential family is determined by its mean and a constant dispersion parameter, indicating the specific distribution used.

Table 3.
Shows the distribution of GMLs with their link function.

Distribution	Link Function	Variance Function
Gaussian	Identity	ϕ
Binomial	Logit	$\frac{\mu_i(1-\mu_i)}{\eta_i}$
Poisson	Log	μ_i
Inverse Gaussian	Inverse Square	$\phi{\mu_i}^3$
Gamma	Inverse	$\phi{\mu_i}^2$

3.4. Random Forest

Breiman [20] introduced the Random Forest (RF) classifier, which is an ensemble learning method built on multiple classification trees $h_k(X|\theta_k)$, where each tree has parameters θ_k randomly picked from a model random vector Θ.

For the final classification, the RF algorithm aggregates the predictions from all trees in the ensemble. Specifically, given an input X each tree $h_k(X)$ cast a vote for the class label, and the class with the most popular votes wins.

Formally, provided dataset $D = \{(x_i, y_i)\}_{i=1}^{n}$, we train multiple classifier $h_k(X)$, where each classifier is defined as:

$$h_k(X) = h(X|\theta_k) \tag{7}$$

Each tree predictor is associated with an outcome $y \in \{\pm 1\}$ for classification tasks.

For regression the algorithm is given:

(a) From the training data, create a bootstrap sample **Z∗** of size N.

(b) To the bootstrapped data, grow a random forest tree T_s by recursively repeating the following procedures for each terminal node of the tree until the minimal node size n_{min} is attained.

 I. From the p variables, choose m variables at random.

 II. Choose the optimal variable/split point from the m.

 III. Divide the node into two daughters.

1. Ensemble on output $\{T_s\}_1^s$ trees

Now predict new variable x given by:

$$f_{r_f}^{S}(x) = \frac{1}{S}\sum_{s=1}^{S} T_s(x) \tag{8}$$

This approach ensures robustness and reduces variance while maintaining high predictive accuracy

3.5. XGBoost

XGBoost is a method that includes Friedman [21] boosting model which was created by Chen and Guestrin [22]. It is a boosting integration model that combines the gradient boost method and decision trees. Instead of utilising the search method, XGBoost directly uses the loss function's first and second derivative values, improving algorithm performance through approaches such as pre-ordering and node number of bits. After incorporating the regularisation term, the XGBoost model selects a basic model that performs well. The regularisation item is utilised in each iteration to reduce weak learner overfitting and does not contribute to the final model's integration. Taylor's formula is applied in each iteration to enlarge the goal function [22]. The equation is:

$$s^{(t)} = \sum_{t=1}^{n}[l\left(y_t, \hat{y}^{(t-1)}\right) + z_i f_t(x_i) + \frac{1}{2}h_i f_t(x)^2] + \varphi(f_t) \qquad (9)$$

Where $\widehat{y}^{(t-1)}$ represents the predictive values of the $(t-1)^{th}$iterations, z_i and h_i are the first and second derivates, $\varphi(f_t)$ is the regulation item, t is the t^{th} interaction, and i is the i^{th}sample. The complexity of the tree is given by:

$$\varphi = \gamma^{T_t} + \frac{1}{2}\sum_{j=1}^{T} \omega_j^2 \qquad (10)$$

Where T_t is the number of leaf nodes in the round t iteration, ω_j represents the weight of the j^{th}leaf node.

3.6. Gradient Boost Model

The GBM is a boosting-like algorithm for regression [21]. The algorithm iteratively combines weak learners with those that do marginally better than the RF to create strong learners [23]. For a given training data set $\delta = \left\{x_{i_3} y_i\right\}_i^n$, the aim of the algorithm is to estimate an approximation $\omega(\boldsymbol{x})$, in the function $\omega(\boldsymbol{x})$, which directs the input function X into the output y, by reducing the predicted value of the given loss function, $L\left(y_1, \omega(\boldsymbol{x})\right)$. Then the algorithm builds an addictive approximation $\omega * (\boldsymbol{x})$ as the sum of weighted functions

$$\omega_m(x) = \omega_{m-1}(x) + p_m h_m(x) \qquad (11)$$

Where the weight of the m^{th} function, $h_m(x)$ is p_m.

These functions represent the ensemble's models, such as decision trees. The approximation is created iteratively. First, a constant approximation of $\omega * (\boldsymbol{x})$ is derived as

$$F_0(x) = arg\,min \sum_{i=1}^{N} L(y_i, \alpha) \qquad (12)$$

The following models are predicted to minimise:

$$(p_{m,} h_m(\boldsymbol{x})) = arg\,min\ \sum_{i=1}^{N} L\left(y_i, \omega_{m-1}(\boldsymbol{x_i})\right) + ph(\boldsymbol{x_i})) \qquad (13)$$

Instead of explicitly addressing the optimisation issue, consider each h_m as a greedy step in a gradient descent optimisation for $\omega\ *$. To train each model, hm, on a fresh dataset $\delta = \left\{x_{i_3} y_i\right\}_i^n$, the pseudo-residuals, r_{mi}, are computed using:

$$r_{mi} = \left[\frac{\partial L\left(y_i, \omega(x)\right)}{\partial \omega(x)}\right]_{\omega(\boldsymbol{x}) = \omega_{m-1}(x)} \qquad (14)$$

The value of p_m is determined by solving a line search optimisation issue. If the iterative procedure is not correctly regularised, this technique may experience overfitting . If the model completely fits the pseudo-residues for certain loss functions, such as quadratic loss, the process may end prematurely if the pseudo-residues become zero in the subsequent iteration. Multiple regularisation hyper-parameters are used to manage the additive process of gradient boosting.

Several regularisation hyper-parameters are used to manage the additive gradient boosting process. To regularise gradient boosting, use shrinkage to lower each gradient decent step: $\omega_m(\boldsymbol{x}) = \omega_{m-1}(\boldsymbol{x}) + v p_m h_m(\boldsymbol{x})$, where v = (0, 1.0]. The value of v is typically set to 0.1. Limiting the complexity of learned models allows for additional regularisation. To restrict the depth of decision trees, we can provide the minimum number of instances required to divide a node. Unlike random forest, gradient boosting's default hyper-parameters limit the expressive capability of trees (e.g., depth to $\approx 3 - 5$) [24]. Finally, another family of hyper-parameters is provided in the various versions. Gradient boosting techniques, such as random subsampling, can enhance ensemble generalisation [21].

3.7. Artificial Neural Network

ANN is a widely used supervised machine learning technique among a variety of domains. It comprises three layers: input, hidden, and output. Its performance is heavily influenced by its structure, including the hidden layers and neurons used [25]. In this study, the feedforward back-propagation neural network, a multilayer perceptron network, is chosen for its straightforward methodology and broad application. Backpropagation (BP) is a very efficient and widely used learning method in multi-layer networks [26, 27].

Now suppose there are n neurons in the input layer, seven in the hidden layer, and two in the output layer. Let the input vector be:

$$X_k = \left(x_{1k} + x_{2k,\dots,} x_{nk}\right) \text{ for } k = 1, 2, \dots, m. \qquad (15)$$

The weights α_{ij} link the input layer to the hidden layer, while β_{ij} connect the hidden layer to the output layer value, with i = 1, 2,..., n, j = 1, 2,..., 5, and l = 1, 2. Neurons within the same layer are not interconnected, but connections exist between the input, hidden, and output layers.

Assuming the activation function is the sigmoid function, input samples $X_1; X_{2,\dots,} X_m$ are sequentially processed. Selecting the k-th input sample X_k , the corresponding hidden layer input vector is:

$$Y_k = \left(y_{1k} + y_{2k,\dots,} y_{7k}\right) \qquad (16)$$

And the hidden layer output is:

$$Z_k = \left(z_{1k} + z_{2k,\dots,} z_{nk}\right) \qquad (17)$$

The output vector is denoted as:

$$\widehat{\mathrm{Y}}_{\mathrm{k}} = (\hat{y}_{1k}, \hat{y}_{2k}) \qquad (18)$$

And the output layer output as:

$$\hat{\mathrm{Z}}_{\mathrm{k}} = (\hat{z}_{1k}, \hat{z}_{2k}) \tag{19}$$

The threshold of each neuron in the output layer is denoted as a_j and b_j the threshold of each neuron in the output layer the inspirit function is given by $f(\varphi_i)$, μ the learning parameter and expected is given by:

$$\frac{1}{2}\sum_{i=1}^{2}(r_{lk} - \hat{z}_{lk})^2 \tag{20}$$

To compute the input and output of neurons in the hidden and output layers, the following derivatives are used:

$$y_{jk} = \sum_{j=1}^{n} \alpha_{ij} x_{ik} - a_j, \tag{21}$$

$$z_{jk} = f\left(y_{j_k}\right), \tag{22}$$

$$\hat{y}_{1k} = \sum_{j=1}^{7} \beta_{jl}\, z_{jk} - b_l, \tag{23}$$

$$\hat{Z}_{lk} = f(\tilde{y}_{lk}) \tag{24}$$

To calculate the partial derivative of the error function for each neuron in the output layer, the predicted and actual network outputs are derived as follows.

$$\frac{\partial E}{\partial \hat{y}_{lk}} = \frac{\partial\,[\frac{1}{2}\sum_{l=1}^{2}(r_{lk} - \hat{z}_{lk})^2]}{\partial \hat{y}_{lk}}$$

$$= \frac{\partial\,[\frac{1}{2}\sum_{l=1}^{2}\left(r_{lk} - f(\tilde{y}_{lk})\right)^2\,]}{\partial \hat{y}_{lk}}$$

$$= -\sum_{1=1}^{2}(r_{lk} - \hat{z}_{lk}) f^1(\hat{y}_1) \underset{=}{\overset{\Delta}{-}} - \delta_{jk} \tag{25}$$

To compute the partial derivative of the error function for each neuron in the hidden layer, the algorithm uses the output layer, output layer, and output of the hidden layer as follows:

$$\frac{\partial E}{\partial y_{lk}} = \frac{\partial\,[\frac{1}{2}\sum_{l=1}^{2}(r_{lk} - \hat{z}_{lk})^2]}{\partial y_{lk}} \quad = -\left(\sum_{1=1}^{2} \delta_{lk}\beta_{jl}\right) f'(y_{jk}) \quad = -\mathrm{p}_{\mathrm{jk}}. \tag{26}$$

Using the above two formulae, we can calculate the change in weight value (β_{jl}) for each modification by:

$$\Delta\beta_{j_l} = -\mu \frac{\partial E}{\partial \beta_{ji}} \cdot \frac{\partial \tilde{y}_{ik}}{\partial \beta_{jl}} = \mu \delta_{1k} z_{jk}. \tag{27}$$

Immediately follows N adjustments, the $(N+1)^{th}$ value is a

$$\beta_{j_l}^{N+1} = \beta_{ji}^{N} + \Delta\beta_{j_l} \tag{28}$$

Similarly, we can obtain the change in weight value α_{ij} in each adjustment and the (N + 1) th value after N adjustments.

$$\Delta\alpha_{ij} = -\mu \frac{\partial E}{\partial \alpha_{ij}} = \mu x_{ik} P_{jk} \tag{29}$$

$$\alpha_{ij}^{N+1} = \alpha_{ij}^{N} + \Delta\alpha_{ij}, \tag{30}$$

The global error can be calculated as follows.

$$E = \frac{1}{2m}\sum_{k=1}^{m} \sum_{i=1}^{2}(r_{lk} - \hat{z}_{lk})^2\,. \tag{31}$$

Finally, in the algorithm, we compare the size of the global error with the setting error. If the global error exceeds the setting error, we keep adjusting the weights until the setting error is met.

3.8. Risk Evaluation and Model Assessment

Evaluating the performance of the crossbreed models , RMSE, R^2 and MAE are used. The mathematical formulas for these metrics are explained below.

3.9. RMSE

RMSE is the average difference between a statistical model's projected values and its actual results. It is mathematically defined as the residuals' standard deviation. The residuals reflect an average distance between the regression line and the data points.

$$RMSE = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(\bar{Y}_j - Y_j\right)^2} \tag{32}$$

3.10. MAE

MAE is a measure of the average magnitude of errors in a set of predictions, without regard for direction [28]. It is calculated as the average absolute difference between predicted and actual values and used to evaluate the efficacy of a regression model.

$$MAE = \frac{1}{N}\sum_{j=1}^{N} |\, \bar{Y}_j - Y_j \,|^2 \tag{33}$$

R^2

The R^2 approach is used to forecast and explain a model's future results [29]. This method is sometimes referred to as R squared and serves as a guideline for assessing the model's correctness.

$$R^2 = 1 - \frac{\sum_{h=1}^{N}(\hat{Y}_h - \bar{Y}_h)^2}{\sum_{n=1}^{N}(Y_h - \bar{Y}_h)^2} \qquad (34)$$

4. Results

4.1. Comparative Analysis

The results in Table 4 evaluate the standard method (GLM with Gamma distribution) and the optimised iterative (hybrid) method demonstrates a substantial improvement in predictive performance when using machine learning approaches such as XGBoost, RF, GBM, and ANN.

On the training set, the standard GLM achieved a modest coefficient of determination ($R^2 = 0.23$), alongside relatively high error values (RMSE = 83,992; MAE = 56,943). In contrast, the hybrid machine learning models yielded markedly higher explanatory power, with R^2 values consistently around 0.39–0.40, and lower error metrics. Notably, XGBoost and ANN exhibited the best balance of predictive fit, producing the lowest RMSE (≈ 82,066 and 82,078, respectively) relative to other models.

On the test set, the performance gap was even more evident. The GLM baseline recorded $R^2 = 0.19$ with higher errors (RMSE = 84,789; MAE = 60,897), whereas all hybrid models outperformed it substantially. XGBoost, RF, and ANN each achieved $R^2 = 0.42$, with RMSE values near 81,200 and MAE values close to 61,500. These results confirm the robustness of the machine learning approaches, which not only generalised better than the standard method but also delivered lower prediction errors.

Overall, the findings suggest that the optimised iterative hybrid method is superior to the standard GLM baseline in both training and testing phases. Among the tested models, XGBoost, RF, and ANN demonstrate the strongest generalisation performance, making them more suitable for practical predictive applications in this context.

Table 4.
Optimised iterative methods against the GLM with GAMMA distribution

Standard Method	**Optimised iterative method (Hybrid)**	R^2	**RMSE**	**MAE**
GLM(GAMMA) Train	NONE	0.23	83992	56943
	XGBoost	0.40	82066	63599
	RF	0.40	82434	63821
	GBM	0.39	83033	64031
	ANN	0.40	82078	63600
GLM(GAMMA) Test	None	0.19	84789	60897
	XGBoost	0.42	81200	61500
	RF	0.42	81222	61589
	GBM	0.40	83032	62300
	ANN	0.42	81221	61549

4.2. Test Statistics

H_0: The standard GLM with a Gamma distribution and the optimised iterative hybrid approach yield similar predictive performance.

H_a: The converse is true

α : 0.05

Table 5.
Two-Sample t-Test Results for Model Performance Measures.

Performance Measures	**Method**	**Variances**	**DF**	**T Value**	**Pr > |t|**
R^2	Pooled	Equal	8	-3.33	0.0104
	Satterthwaite	Unequal	3.5651	-5.03	0.0099
MAE	Pooled	Equal	8	16.61	<.0001
	Satterthwaite	Unequal	1.0898	9.6	0.055
RMSE	Pooled	Equal	8	5.03	0.001
	Satterthwaite	Unequal	3.0798	7.26	0.0049

5. Conclusion

At the 5% significance level, the optimised iterative hybrid GLM demonstrates statistically significant differences from the standard GLM across key predictive performance measures. The improvement in model fit is supported by significantly higher R^2 values under both variance assumptions. RMSE is also consistently and significantly different, indicating a meaningful change in predictive precision. MAE shows a significant difference under the pooled assumption but not under the unequal variance test, suggesting that this result should be interpreted with caution. Taken together, these findings indicate that the optimised GLM offers measurable improvements in explanatory power, although the evidence regarding predictive accuracy is more mixed.

6. Discussion

The comparative analysis between the standard Generalised Linear Model (GLM) with a Gamma distribution and the optimised iterative hybrid method highlights the advantages of employing advanced machine learning techniques in predictive modelling. The baseline GLM demonstrated relatively weak explanatory power ($R^2 = 0.23$ on training, 0.19 on testing) alongside comparatively higher prediction errors (RMSE = 83,992–84,789; MAE = 56,943–60,897). In contrast, hybrid machine learning models XGBoost, RF, GBM, and ANN consistently achieved superior performance.

On both the training and testing datasets, XGBoost, RF, and ANN delivered the strongest predictive performance, each producing R^2 values around 0.40–0.42 and lower error metrics relative to the GLM baseline. These results confirm that the optimised hybrid method generalises better to unseen data, thereby reducing the risk of overfitting while maintaining predictive accuracy. XGBoost and ANN provided the lowest RMSE scores, underscoring their capacity to minimise prediction error. This aligns with the findings of previous studies, such as those by Buthelezi et al. (2024) [4] and Wilson et al. (2024) [6], which highlight the potential of ML techniques to outperform conventional actuarial models in predictive tasks.

The formal hypothesis testing further substantiates these empirical observations. At a 5% significance level, the hybrid models exhibited statistically significant improvements in R^2 and RMSE, reinforcing the claim that they offer enhanced explanatory power and predictive precision compared with the standard GLM. While the results for MAE were significant under the pooled variance assumption, they did not hold under the unequal variance test, suggesting that improvements in absolute error should be interpreted with caution. This nuance highlights that while the hybrid models generally outperform the GLM, the magnitude of improvement may vary depending on the performance metric considered.

7. Conclusion

The findings of this study demonstrate that the optimised iterative hybrid approach offers a robust and statistically significant improvement over the standard GLM with Gamma distribution. The machine learning models particularly XGBoost, RF, and ANN not only enhanced explanatory power but also improved predictive precision, as evidenced by higher R^2 values and lower RMSE scores across both training and testing phases.

Although the evidence for improvements in MAE is less conclusive, the overall performance gains indicate that the hybrid method represents a more effective predictive modelling framework than the traditional GLM baseline. Consequently, these results suggest that integrating iterative optimisation with advanced ML algorithms can provide a more reliable and generalisable solution for predictive analytics in this context.

7.1. Limitations

Despite the promising results, this study has few limitations that warrant consideration. Firstly, the analysis is based on historical claims data from a single South African non-life vehicle insurance company, which may limit the generalisability of the findings to other contexts or regions. Future research should consider a broader dataset encompassing diverse geographical and market conditions to validate the robustness of the hybrid models. Furthermore, the study employs specific ML techniques and hyperparameter tuning methods, which may not represent the full spectrum of available optimisation strategies. Exploring alternative ML algorithms and optimisation techniques could provide additional insights into the most effective approaches for enhancing GLM performance in insurance pricing.

References

[1] F. Greberg and A. Rylander, "Using gradient boosting to identify pricing errors in GLM-based tariffs for non-life insurance," Master’s Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2022.

[2] S. Xie and K. Shi, "Generalised additive modelling of auto insurance data with territory design: A rate regulation perspective," *Mathematics,* vol. 11, no. 2, p. 334, 2023. https://doi.org/10.3390/math11020334

[3] Y. Havrylenko and J. Heger, "Detection of interacting variables for generalized linear models via neural networks," *European Actuarial Journal,* vol. 14, no. 2, pp. 551-580, 2024. https://doi.org/10.1007/s13385-023-00362-4

[4] S. J. Buthelezi, T. Hungwe, S. M. Seeletse, and V. Mbirimi-Hungwe, "Non-life insurance: The state of the art of determining the superior method for pricing automobile insurance premiums using archival technique," *International Journal of Research in Business and Social Science,* vol. 13, no. 2, pp. 180-188, 2024. https://doi.org/10.20525/ijrbs.v1312.3211

[5] V. Hassija *et al.*, "Interpreting black-box models: A review on explainable artificial intelligence," *Cognitive Computation,* vol. 16, no. 1, pp. 45-74, 2024. https://doi.org/10.1007/s12559-023-10179-8

[6] A. A. Wilson, A. Nehme, A. Dhyani, and K. Mahbub, "A comparison of generalised linear modelling with machine learning approaches for predicting loss cost in motor insurance," *Risks,* vol. 12, no. 4, p. 62, 2024. https://doi.org/10.3390/risks12040062

[7] P. Panjee and S. Amornsawadwatana, "A generalized linear model and machine learning approach for predicting the frequency and severity of Cargo insurance in Thailand’s border trade context," *Risks,* vol. 12, no. 2, p. 25, 2024. https://doi.org/10.3390/risks12020025

[8] C. Clemente, G. R. Guerreiro, and J. M. Bravo, "Modelling motor insurance claim frequency and severity using gradient boosting," *Risks,* vol. 11, no. 9, p. 163, 2023. https://doi.org/10.3390/risks11090163

[9] F. Holvoet, K. Antonio, and R. Henckaerts, "Neural networks for insurance pricing with frequency and severity data: A benchmark study from data preprocessing to technical tariff," *North American Actuarial Journal,* pp. 1-44, 2025. https://doi.org/10.1080/10920277.2025.2451860

[10] S. Ardabili, A. Mosavi, and A. Várkonyi-Kóczy, *Advances in machine learning modeling reviewing hybrid and ensemble methods. InInternational conference on global research and education 2019 Sep 4*. Cham: Springer International Publishing, 2019.

[11] A. Brauer, "Enhancing actuarial non-life pricing models via Transformers," *European Actuarial Journal,* vol. 14, no. 3, pp. 991-1012, 2024. https://doi.org/10.1007/s13385-024-00388-2

[12] R. Van Oirbeek, F. Vandervorst, T. Bury, G. Willame, C. Grumiau, and T. Verdonck, "Non-differentiable loss function optimization and interaction effect discovery in insurance pricing using the genetic algorithm," *Risks,* vol. 12, no. 5, p. 79, 2024. https://doi.org/10.3390/risks12050079

[13] F. Novkaniza, I. D. Putri, R. Al Kafi, and S. Devila, "A posteriori premium rate calculation using poisson-gamma hierarchical generalized linear model for vehicle insurance," *Jurnal Teori dan Aplikasi Matematika,* vol. 9, no. 1, pp. 221-241, 2025.

[14] E. Seyam and H. Elsalmouny, "Proposed models for comprehensive automobile insurance ratemaking in Egypt with parametric and semi-parametric regression: a case study," *Journal of Statistics Applications & Probability,* vol. 11, no. 1, pp. 41-55, 2022.

[15] L. Berry, "Hybrid hidden Markov model and generalized linear model for auto insurance premiums," Doctoral Dissertation, Concordia University, 2016.

[16] J. P. C. Reil, "Beyond generalized linear models: advancing insurance pricing through interpretable and explainable machine learning," Master's Thesis, University of Twente, 2024.

[17] C. Arnold, L. Biedebach, A. Küpfer, and M. Neunhoeffer, "The role of hyperparameters in machine learning models and how to tune them," *Political Science Research and Methods,* vol. 12, no. 4, pp. 841-848, 2024. https://doi.org/10.1017/psrm.2023.61

[18] X. Bouthillier, C. Laurent, and P. Vincent, "Unreproducible research is reproducible," presented at the InInternational Conference on Machine Learning 2019 May 24 (pp. 725-734). PMLR, 2019.

[19] J. A. Nelder and R. W. M. Wedderburn, "Generalized linear models," *Journal of the Royal Statistical Society, Series A (Statistics in Society),* vol. 135, no. 3, pp. 370–384, 1972. https://doi.org/10.2307/2344614

[20] L. Breiman, "Random forests," *Machine learning,* vol. 45, no. 1, pp. 5-32, 2001.

[21] J. Friedman, "Greedy function approximation: A gradient boosting machine," *Annals of Statistics,* vol. 29, pp. 1189-1232, 2001.

[22] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). San Francisco, CA, USA: ACM*, 2016.

[23] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence,* vol. 14, no. 771-780, p. 1612, 1999.

[24] C. Bentéjac, A. Csörgő, and G. Martínez-Muñoz, "A comparative analysis of gradient boosting algorithms," *Artificial Intelligence Review,* vol. 54, no. 3, pp. 1937-1967, 2021. https://doi.org/10.1007/s10462-020-09896-5

[25] M. Chester, *Neural networks: A tutorial*. Englewood Cliffs, NJ, USA: PTR Prentice Hall, 1993.

[26] M. Hajihassani, D. J. Armaghani, H. Sohaei, E. T. Mohamad, and A. Marto, "Prediction of airblast-overpressure induced by blasting using a hybrid artificial neural network and particle swarm optimization," *Applied Acoustics,* vol. 80, pp. 57-67, 2014. https://doi.org/10.1016/j.apacoust.2014.01.005

[27] W. Yu *et al.*, "Claim amount forecasting and pricing of automobile insurance based on the BP neural network," *Complexity,* vol. 2021, no. 1, p. 6616121, 2021. https://doi.org/10.1155/2021/6616121Digital

[28] T. Chai and R. R. Draxler, "Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature," *Geoscientific Model Development,* vol. 7, no. 3, pp. 1247-1250, 2014.

[29] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *Peerj Computer Science,* vol. 7, p. e623, 2021. https://doi.org/10.7717/peerj-cs.623