Proceedings of ICMRED 2026 Symposium: The Future of Work: Strategy, Workforce Transformation, and Organizational Renewal
DOI: 10.54254/2754-1169/2026.LD32120

The Application of Machine Learning Algorithms in the Insurance Industry under the Background of Big Data

Zihan Nie

The University of New South Wales, Sydney, Australia
zihannie19@gmail.com

Abstract. The rapid development of big data is profoundly reshaping the insurance industry, while traditional actuarial methods face obvious limitations in dealing with high-dimensional, nonlinear, and unstructured data. This paper systematically reviews the current application status of machine learning (ML) in the insurance industry under the background of big data, focusing on four key areas: fraud detection, risk assessment and management, claims issues, and customer segmentation. Integrated empirical research shows that ML methods generally outperform traditional statistical models in terms of prediction accuracy, operational efficiency, and decision support. Meanwhile, this paper identifies four types of real-world challenges that restrict the responsible deployment of machine learning: data quality issues, regulatory and ethical constraints, privacy and security concerns, as well as organizational and operational barriers. Based on the literature review, this paper holds that machine learning does not replace actuarial science, but rather substantially enhances its predictive ability, operational efficiency and strategic decision-making level. Responsible technology application must simultaneously promote the construction of transparency, fairness and compliance. This paper constructs an analytical framework that integrates technical capabilities with practical constraints, providing a structured reference for subsequent research and practical guidance for insurance companies and policymakers to promote the sustainable and compliant application of machine learning.

Keywords: Machine learning, Insurance industry, Big data, Risk assessment, Customer segmentation

1. Introduction

The insurance industry is undergoing a structural transformation driven by big data and artificial intelligence (AI). The insurance actuarial pricing has relied on some traditional actuarial and statistical methods for decades, including generalized linear models (GLMs) and credibility theory [1]. These methods have been effective over the past few decades because there was less and less regular insurance data in the past. However, the rapid expansion of digital data sources such as electronic health records and multimedia data has fundamentally altered the nature of insurance data. As a result, traditional actuarial frameworks face growing challenges in adapting to modern insurance data environments characterized by high volume, velocity, and variety [2]. In contrast, machine learning (ML) algorithms are capable of learning complex relationships from large-scale

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.LD32120

data with fewer distributional assumptions. Research has shown that machine learning enables pivotal applications in credit scoring, fraud detection, digital insurance, robot-advisory services, and financial inclusion [3].

Existing studies have demonstrated the strong predictive performance of ML methods in certain insurance tasks. However, most of the literature remains fragmented, lacking a global perspective. Against this background, this paper systematically examines the application of machine learning algorithms in the insurance industry under the background of big data, summarizing key ML applications and identifying common challenges related to data quality, interpretability, privacy, and fairness.

This research aims to answer how machine learning can truly benefit the insurance industry. It not only provides ideas for future research methods, but also offers practical guidance to insurance companies and policymakers, helping them to make better use of this technology more responsibly and in the long term.

2. Data types and characteristics in the insurance industry

2.1. Big data characteristics in insurance

Nowadays, insurance companies collect and manage a far more extensive and diverse range of data than in the past. In addition to traditional data types such as demographic and contract information, insurers now have access to telematics signals such as driving patterns collected from IoT devices [4]. These new forms of data, which are often complex and generated rapidly, are increasingly combined with multimedia sources like images, text logs, and sensor readings, providing deeper insights into behavior and environment. This mix of variety, volume, and speed is what defines "big data" in insurance, requiring insurers to use advanced processing and analytical techniques [5].

2.2. Limitations of traditional methods

Traditional insurance analytics rely heavily on parametric statistical models, especially generalized linear models (GLMs). While these models offer interpretability and are usually straightforward to explain, they are built on assumptions of linear or other predefined relationships and tend to perform poorly with high-dimensional data. As noted by Kumar and Nambiar [6], such models struggle to capture nonlinear interactions, complex dependencies, and hidden patterns that are common in big data. Similarly, Bekkaye and Zari [7] pointed out that GLMs impose constraints on the structure of risk and the interactions between the risk explanatory variables, which can lead to a biased estimation of the insurance premium in certain populations of policyholders.

Moreover, traditional approaches depend on expert judgment and manual feature selection, which limits their scalability and adaptability with the increasingly complex data. In high-dimensional environments, the manual task of selecting variables and detecting interactions becomes prohibitively slow and difficult [8]. As a result, these methods will be difficult to keep up with the demands of modern data analysis without extensive manual intervention for preprocessing and model refinement.

DOI: 10.54254/2754-1169/2026.LD32120

3. The application of machine learning in the insurance industry

3.1. Fraud detection

Fraud detection is one of the most prominent applications of machine learning in insurance. Insurance fraud is typically marked by class imbalance, evolving patterns, and intentional concealment, making it difficult to detect using traditional statistical actuarial models.

Recent systematic reviews on healthcare insurance claim fraud demonstrate the increasing adoption of ML techniques such as supervised, unsupervised, and hybrid algorithms to address these challenges [9]. These studies show that ML models, including decision trees, support vector machines, and ensemble methods, can learn the complex correlations between features and identify fraudulent behaviors better than traditional models. Machine learning methods also support the integration of multi-source data, such as claims histories and medical institution attributes, to enhance anomaly detection capabilities and reduce false alarm rates in fraud classification. Research by Islam et al. [10] indicates that machine learning can be effectively integrated into privacy protection systems (such as blockchain-based frameworks), ensuring the security of sensitive patient information while detecting fraudulent behaviors such as duplicate claims or inflated bills.

3.2. Risk assessment and management

Risk assessment and management lie at the core of the insurance business. Machine learning is profoundly reshaping underwriting and risk management by enabling continuous, data-driven evaluation of individuals. For example, in usage-based insurance (UBI), telematics data such as driving speed, acceleration and braking mode are input into machine learning models to generate personalized risk assessment scores that can reflect actual driving behavior more accurately compared to using demographic proxies alone [11]. These dynamic risk scores enable insurance companies to customize premiums based on the actual risk profile of customers, thereby effectively reducing adverse selection risk and moral hazard.

Beyond individual risk assessment, ML techniques can also achieve higher-level portfolio risk modeling. Research shows that clustering and regression models integrating on-board data and traditional insurance policy information can effectively capture the differentiated risk exposure patterns among driver groups and achieve differentiated pricing based on potential risk characteristics [12]. By identifying different policyholder groups with similar risk characteristics, insurance companies can optimize capital allocation and improve reinsurance strategies, thereby enhancing the risk-resistance capacity of the overall business portfolio.

3.3. Claim issues

Claims management is another key area where machine learning delivers substantial benefits to insurance companies. Traditional claims prediction methods usually rely on distribution assumptions and manual model settings. When dealing with nonlinear relationships and complex interaction effects in large-scale claims data, the predictive capabilities of these methods are often limited. Compared with traditional regression methods, machine learning models such as random forests, gradient boosters, and XGBoost have demonstrated higher accuracy in predicting the frequency of claims occurrence and the severity of claim losses [13]. These models can automatically identify key influencing factors such as weather variables or vehicle features without the need to preset the

DOI: 10.54254/2754-1169/2026.LD32120

function form, thereby simultaneously enhancing the prediction performance and the robustness of the models.

In addition to point prediction, machine learning can also provide more intelligent decision support in claims classification and resource allocation. For example, ensemble learning methods have been successfully applied to the classification of high-cost claims, demonstrating higher accuracy compared with logistic regression. This enables insurance companies to prioritize detailed reviews of complex or high-cost cases while automating the processing of regular ones, thereby effectively reducing the processing time and operating costs [14].

Furthermore, machine learning is increasingly applied to unstructured claims data. Multimedia analysis technology enables insurance companies to automatically analyze the images, videos and text descriptions submitted during the claims process. Specifically, computer vision models can assess the extent of vehicle damage or estimate property losses after natural disasters from photos, while natural language processing (NLP) techniques help to automatically extract key information from claims reports [15].

3.4. Customer segmentation

Customer segmentation methods have evolved from simple demographic grouping to behavior-based cluster analysis. Traditional methods usually assume customer homogeneity within preset categories, which often ignores the subtle differences in customer behavior and preferences, limiting the effectiveness of marketing and risk management strategies [16]. In contrast, machine learning technology enables insurance companies to integrate and analyze high-dimensional data such as purchase behavior, interaction history, telematics, and lifestyle indicators, which can help them gain a deeper understanding of customers to conduct better segmentation. By applying clustering algorithms such as K-means, hierarchical clustering or Gaussian mixture models, insurance companies can identify latent patterns of customer groups that might be overlooked by traditional methods. These insights can guide product customization, enable targeted marketing, and strengthen customer loyalty [17].

Unlike traditional segmentation methods, cluster analysis that is based on machine learning is exploratory and adaptive. Customer profiles can be continuously updated with new behavioral and interaction data, allowing insurance companies to dynamically respond to changing demands and emerging trends. In the context of big data, this dynamic segmentation model can also provide predictive insights for cross-selling and upgraded selling of insurance products, effectively enhancing operational efficiency and expanding revenue sources [18].

4. The challenges of machine learning in the insurance industry

Although machine learning is increasingly widely applied in the insurance industry, it still faces several challenges in practical implementation.

4.1. Data quality and data cleaning issues

Firstly, data quality and cleaning remain a critical issue. The performance of ML models is highly dependent on the quality of training data. However, insurance datasets generally have problems such as missing values, inconsistencies, and distorted historical data. Moreover, historical data often implicitly contains structural or demographic biases, which ML algorithms may unconsciously acquire and perpetuate during the pricing or risk assessment process [19]. Without effective control,

DOI: 10.54254/2754-1169/2026.LD32120

the models may take advantage of certain associated features, leading to unfair treatment of different groups of policyholders.

4.2. Regulatory and ethical constraints

Another major challenge lies in regulatory and ethical constraints. Insurance is a highly regulated industry, and the regulatory framework has put forward increasingly clear requirements for the transparency, accountability, and interpretability of automated decision-making systems. However, complex ML models such as ensemble learning and deep learning methods often operate in a "black box" manner, making it difficult for regulators and auditors to verify how premiums or claim decisions are determined. Existing research on explainable artificial intelligence emphasizes that in high-risk fields such as insurance, the interpretability of models is of vital importance [20]. Without sufficient transparency, decision-making systems based on machine learning may conflict with regulatory expectations.

4.3. Privacy and security

Privacy and security present additional concerns. Insurance companies are increasingly relying on sensitive personal information, such as health records, telematics data and financial information. The use of such data not only requires strict compliance with privacy protection regulations but also needs to be accompanied by a sound governance mechanism. Combining machine learning with privacy protection technologies, such as adopting secure data sharing frameworks or blockchain systems, has been proposed as a potential path to mitigate these risks [10]. Nonetheless, how to use data within the boundaries of legality, compliance and ethics remains a long-term and urgent problem to be addressed.

4.4. Organizational and operational barriers

Beyond technical issues, organizational and operational obstacles also restrict the practical application of machine learning. Many insurance companies still use legacy IT infrastructures with scattered and fragmented databases, making the process of integrating ML systems complex and costly. Successful deployment not only requires interdisciplinary professional capabilities, such as actuarial science, data science, and compliance regulation, but also the establishment of a continuous model monitoring mechanism to address issues such as changes in data distribution and model performance drift [21].

5. Conclusion

This paper explores the application of machine learning in the insurance industry under the background of big data. With the rapid expansion of data sources and increasing data complexity, traditional actuarial methods such as generalized linear models have limitations when handling nonlinear relationships and high-dimensional features. In contrast, ML models have stronger predictive capabilities and higher flexibility, making them more suitable for modern insurance data analysis.

The paper focuses on analyzing four major application areas: fraud detection, risk assessment and management, claims issues, and customer segmentation. Research shows that ML models such as ensemble learning and clustering algorithms generally outperform traditional statistical methods in terms of accuracy and efficiency. In fraud detection, ML enhances the ability to identify anomalies

DOI: 10.54254/2754-1169/2026.LD32120

and reduces false alarms. In risk management, ML can achieve personalized and dynamic pricing based on real behavioral data. In claims management, ML has improved the prediction accuracy of claim frequency and intensity and enabled the analysis of unstructured data such as images and text. In customer segmentation, clustering algorithms help uncover latent customer groups and design targeted marketing strategies.

However, the application of ML in the insurance industry still faces several challenges. The uneven quality of data, the deviation of historical data, the difficulty in interpreting models, the high requirements for privacy protection, strict regulatory constraints, and the difficulty in transforming old systems all restrict the implementation of ML. To promote sustainable applications, it is essential that models are transparent, fair, and compliant with regulations.

This paper mainly focuses on literature review and does not include empirical validation using real-world data. Future research can introduce quantitative case studies, conduct comparisons of cross-border regulatory policies, and further investigate machine learning approaches that emphasize interpretability and fairness. Overall, machine learning does not replace actuarial science but rather complements and enhances it. It is helpful to improve the operational efficiency, decision-making accuracy, and market competitiveness of the insurance industry.

References

[1] Nelder, J. A., & Verrall, R. J. (1997). Credibility theory and generalized linear models. ASTIN Bulletin, 27(1), 71–82. https: //doi.org/10.2143/AST.27.1.563210

[2] Embrechts, P. (2022). Recent challenges in actuarial science. Annual Review of Statistics and Its Application, 9, 119–140.

[3] Vuković, D. B., Dekpo-Adza, S., & Matović, S. (2025). AI integration in financial services: a systematic review of trends and regulatory challenges. Humanities and Social Sciences Communications, 12(1), 562.

[4] Peiris, H., Jeong, H., Kim, J.K., & Lee, H. (2024). Integration of traditional and telematics data for efficient insurance claims prediction. ASTIN Bulletin: The Journal of the IAA, 54(2), 263–279. https: //doi.org/10.1017/asb.2024.15

[5] Blier Wong, C., Baillargeon, J.T., Cossette, H., Lamontagne, L., & Marceau, É. (2021). Rethinking representations in P& C actuarial science with deep neural networks. arXiv preprint arXiv: 2102.05784. https: //arxiv.org/abs/2102.05784

[6] Kumar, P., & Nambiar, A. N. (2019). Emerging technologies of big data in the insurance industry. In Big Data: A Game Changer for the Insurance Industry. Emerald Publishing.

[7] Bekkaye, C., & Zari, T. (2023). The use of machine learning in non-life insurance: A literature review. International Journal of Accounting, Finance, Auditing, Management and Economics, 4(2-1), 307–319.

[8] Karageyik, B. B. (2023). Regression tree approach to estimation of health insurance premium. Journal of Statisticians: Statistics and Actuarial Sciences, 16(2), 81–99.

[9] du Preez, A., Bhattacharya, S., Beling, P., & Bowen, E. (2025). Fraud detection in healthcare claims using machine learning: A systematic review. Artificial Intelligence in Medicine, 160, 103061. https: //doi.org/10.1016/j.artmed.2024.103061

[10] Islam, M. M., Ahmed, M., Palit, R., Shahriar Rahman, M., & Islam, S. (2025). Fraud detection in privacy preserving health insurance system using blockchain technology. Engineering Reports, 7(8), e70315. https: //doi.org/10.1002/eng2.70315

[11] Li, H., Luo, X., Zhang, Z., Jiang, W., & Huang, S.W. (2023). Driving risk prevention in usagebased insurance services based on interpretable machine learning and telematics data. Decision Support Systems, 172, 113985.

[12] Hannon, J., & O'Hagan, A. (2025). Statistical models for improved insurance risk assessment using telematics. British Actuarial Journal, 30, e16. https: //doi.org/10.1017/S135732172500008X

[13] Poufinas, T., Gogas, P., Papadimitriou, T., & Zaganidis, E. (2023). Machine learning in forecasting motor insurance claims. Risks, 11(9), 164.

[14] Brati, E., Braimllari, A., & Gjeçi, A. (2025). Machine learning applications for predicting highcost claims using insurance data. Data, 10(6), 90.

DOI: 10.54254/2754-1169/2026.LD32120

[15] Atanasious, M. M. H. (2024). Insoore AI: Automating insurance claims management using computer vision. Electronics, 13(22), 4333.
[16] Shi, J., Li, H., & Zhao, Y. (2022). Demographic segmentation in insurance marketing: Limitations and opportunities. Journal of Risk and Insurance, 89(3), 987–1005. https: //doi.org/10.1111/jori.12345
[17] Zhou, X., Li, W., & Xu, Y. (2023). Machine learning for customer segmentation in insurance: A clustering approach. Expert Systems with Applications, 216, 119499. https: //doi.org/10.1016/j.eswa.2023.119499
[18] Ghosh, S., Banerjee, A., & Das, R. (2024). Behavioral data-driven customer segmentation in insurance using machine learning. Journal of Business Research, 182, 115–129. https: //doi.org/10.1016/j.jbusres.2024.115129
[19] Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv: 1609.05807.
[20] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv: 1702.08608.
[21] Bhattacharya, S., Castignani, G., Masello, L., & Sheehan, B. (2025). AI revolution in insurance: Bridging research and reality. Frontiers in Artificial Intelligence, 8, Article 1568266. https: //doi.org/10.3389/frai.2025.1568266