International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

AI-Augmented Risk Prediction in Health Insurance

Sumit Kumar[1], Shailesh Kumar[2], Dushyant Sharma[3]

[1,2,3]Bachelors of Computer Science Chandigarh University, India

Abstract

Risk prediction is central tothe underwriting and pricing processes for health insurers. While traditional actuarial models are built around statistical robustness, they struggle to modelcomplex, non-linear associations between different risk factors. In this paper, wepropose an audience awareness AI-augmented framework for the risk prediction in health insur- ance using the Machine Learning algorithms for improving prediction accuracy. Based on structured health data, electronic healthrecords, and socio-demographic variables, we compare several models: Random Forest, Gradient Boosting, Deep Neural Networks. Ourfindings indicate a marked enhancement in predic- tive performance over conventional logistic regression models and highlight the potential of Artificial Intelligence as a revolutionary asset in the health insurance sector.

Index Terms: Health Insurance, Risk Prediction, Artificial In- telligence, Machine Learning, Deep Learning, Actuarial Science, Generalised Linear Models.

INTRODUCTION

The health industry, along with all of its different aspects, requires risk prediction the most. Risk prediction influences various things such as underwriting, premium calculation, claims forecasting, and policy design. Historically, the in- surers have relied on models based on well-established sta- tistical methods such as logistic regression and generalised linear models [2]. These are interpretable and computation- ally efficient approaches but they still possess several limita- tions—particularly in the face of today's rapidly growing and increasingly complex healthcare datasets.

The emergence of electronic health records (EHRs), wear- able health devices, genomics, and real-time claims data has transformed the healthcare data landscape. These sources generate large volumes of data which is high-dimensional, heterogeneous, and often unstructured. Making use of this highly raw and unstructured data requires the use of models far different than the ones in use already. Moreover, risk factors in healthcare often interact in non-linear and ways specific to a context—making it difficult for linear models to detect underlying patterns without extensive attribute engineering.

Artificial Intelligence (AI) has proven very successful in recognising complex relationships within large datasets across a wide range of domains such as finance, marketing, and most notably, healthcare. By understanding patterns directly from the data, these models can improve "risk-predictive" perfor- mance and adapt to new data more quickly than conventional methods and approaches.

Even after the promises and advancements of AI, the incorporation of AI is still in very early stages in the healthcare industry. Concerns around data privacy and regulatory com- pliance have slowed widespread implementation. But, with recent advancements in explainable AI (XAI) and regulatory

frameworks, the industry is now ready and eager to integrate AI-driven tools more confidently and effectively.

This paper investigates the application of AI to health risk prediction with a focus on health insurance. We compare tradi- tional and modern machine learning models using real-world health and demographic data. We analyse their performance across multiple evaluation metrics and explore their potential to augment or even replace conventional actuarial models. Our study aims to provide a data-driven foundation for insurers seeking to adopt AI for smarter, fairer, and more adaptive risk assessment.

RELATED WORK

Artificial intelligence has been a hot topic in the industry of healthcare for a while now. Yet, the use of AI in health insurance—particularly for predicting risks—hasn't received as much attention, even though industry players are getting increasingly interested.

Predictive modelling was used in healthcare mainly for clinical outcomes like disease onset, mortality risk, etc [6] [**?**]

[8] [10]. For example, Rajkomar et al. applied deep learning techniques to predict various medical outcomes straight from electronic health records (EHRs), showing that they could outperform traditional models [1]. Similarly, Harutyunyan and colleagues delved into multitask learning to predict patient mortality, decompensation, and length of stay all at once, using time-series clinical data [5]. These works underline the effectiveness of machine learning and deep learning models in capturing complex temporal and relational patterns in health- care data.

In the insurance domain, traditional risk assessment models are predominantly actuarial and rely on structured data with assumptions of linearity and independence among variables. These models usually lack the flexibility to utilise unstructured or semi-structured data, such as clinical notes, wearable sensor data, or dynamic behavioural data. Which has resulted in a growing demand for more adaptive and data-driven risk prediction approaches.

Recent work has begun to address this gap. For instance, Kharroubi and Gabrani applied decision tree models and support vector machines (SVMs) to health insurance claims data for classification tasks. Their results suggested that ML models could detect patterns missed by traditional statistical techniques [15]. Another study by Yu and team implemented gradient boosting machines (GBMs) to predict high-cost in- surance claimants, showing improvements in both precision and recall over standard logistic regression models.

Moreover, there is increasing interest in leveraging ex- plainable AI (XAI) tools to enhance the transparency of machine learning models in healthcare insurance applications. Lundberg and Lee introduced SHAP (SHapley Additive ex- Planations), which has become widely adopted for interpret- ing complex models and addressing regulatory requirements regarding fairness and accountability [4] [8].

Despite these advancements, the literature still lacks com- prehensive frameworks that integrate diverse data sources (e.g., demographic, behavioural, clinical, and temporal data) using advanced AI models specifically tailored to health in- surance risk prediction. Most existing studies are limited in scope—focusing either on a specific subset of risk factors or on technical model comparison without considering practical deployment challenges [11].

Our work seeks to bridge this gap by evaluating and comparing multiple AI techniques across a broad set of health risk indicators relevant to insurers. We focus not only on predictive accuracy but also on

interpretability, scalability, and the practical feasibility of deploying these models within insurance organisations. By doing so, this paper contributes to both academic research and industry practice, offering actionable insights into how AI can reshape risk modelling in the health insurance sector.

METHODOLOGY

This study follows a multi-step methodology encompassing data acquisition, preprocessing, feature engineering, model selection, training, and evaluation. The goal is to compare AI-based models with traditional statistical techniques for accurate and explainable risk prediction in health insurance.

A. Data Collection and Sources

The dataset used in this study comprises structured health and demographic data obtained from multiple sources:

Public Health Databases:

Including MIMIC-III and NHANES (National Health and Nutrition Examination Survey), which provide de- identified patient data such as diagnoses, medications, lab results, and hospital records [6] [7].

Synthetic Insurance Data:

Created using a data generator that simulates policyholder behavior (e.g., claims, renewals, coverage types) while reflecting real-world statistical distributions.

Socioeconomic Indicators:

Extracted from public census datasets and linked via ZIP code-level mapping.

Each record includes:

- Personal attributes (age, sex, income, occupation, educa- tion)
- Health history (chronic diseases, number of medications, hospitalisation history)
- Lifestyle (BMI, smoking, alcohol use, exercise fre- quency)
- Claims history (number, amount, frequency, denial rate) Over 100,000 records were aggregated, with features span-

ning both static and temporal domains.

B. Data Preprocessing

Missing Data Handling:

- Categorical variables were imputed using the mode.
- Continuous variables (e.g., BMI, income) were im- puted using mean or median values.
- For time-series gaps (in claims history), forward-fill methods were applied with decay factors.

Normalisation and Encoding:

- Numerical features were standardised using z-score normalisation.
- Categorical variables (e.g., gender, smoker status) were encoded using one-hot encoding.
- Date fields (e.g., policy start, last hospitalisation) were transformed into duration features (e.g., days since event).

Outlier Detection and Treatment:

- Outliers in features like total annual claims and hospital visits were capped at the 99th percentile using winsorisation to mitigate their impact on model training.

C. Feature Engineering

We engineered over 200 features, grouped into five primary categories:

1. Demographic Features: Age, gender, marital status, ed- ucation level, income bracket, and region.

2. Medical History Features: Number of diagnosed chronic illnesses (e.g., diabetes, hypertension), prior surgeries, and total hospital stays in the past 5 years.
3. Behavioural Features: Smoking status, exercise fre- quency, dietary patterns, and alcohol consumption.
4. Claims Features: Number of claims in the past year, average claim amount, frequency of denials, time since last claim.
5. Temporal Features: Rolling averages and trends (e.g., increasing claims over the last 6 months) and time-series encoding (e.g., Fourier transforms for seasonal patterns).

SHAP analysis was used to validate feature importance post- training.

D. Model Architectures and Training

We implemented five models to compare predictive perfor- mance:

Logistic Regression (Baseline)

A standard GLM model used in traditional actuarial science, providing interpretable coefficients.

Random Forest

A robust ensemble method suitable for high-dimensional tabular data. 500 trees were used with Gini impurity as the splitting criterion.

XGBoost

Gradient boosting decision tree model optimised with regularisation to prevent overfitting [3]. Hyperparame- ters were tuned using grid search:

- max depth = 6
- learning rate = 0.1
- n estimators = 100
- subsample = 0.8

Deep Neural Network (DNN):

- Architecture: Input layer → [256 → 128 → 64] hidden layers → output (sigmoid)
- Activation: ReLU
- Optimiser: Adam
- Loss Function: Binary cross-entropy
- Dropout (0.3) used for regularisation.

LSTM (Long Short-Term Memory):

Designed for sequential claims data. Each patient's claim history was encoded as a time-series sequence:

- Input: claim count and cost over 12 monthly inter- vals
- Hidden units: 128
- Return sequences: False
- Final dense layer with sigmoid activation.

All models were trained on 80 percent of the data, with 10 percent used for validation and 10 percent for testing.

E. Evaluation Metrics

Each model was evaluated based on both classification and probabilistic performance:

- Area Under ROC Curve (AUC): Measures discrimination capability.
- F1-Score: Balances precision and recall, especially useful in imbalanced datasets.
- Brier Score: Measures the accuracy of probabilistic pre- dictions.

- Calibration Curve: Assesses alignment between predicted and observed probabilities.
- Confusion Matrix Analysis: Used to analyse false posi- tives/negatives for practical risk categorisation.

A 5-fold cross-validation was applied for all models to ensure generalisability and reduce overfitting.

RESULTS

A. Model performance summary

LSTM outperformed all other models, especially in recognising temporal trends and sequential dependencies in claim behavior.

TABLE I RESULTS

Model	AUC	Accuracy	Precision	Recall	F1-Score	Brier Score
Logistic Regres- sion	0.72	0.69	0.63	0.66	0.65	0.210
Random Forest	0.82	0.77	0.71	0.75	0.73	0.170
XGBoost	0.85	0.79	0.74	0.78	0.76	0.150
DNN	0.87	0.81	0.77	0.80	0.78	0.130
LSTM	0.89	0.83	0.79	0.82	0.80	0.120

- DNN closely followed, benefiting from high-dimensional feature interactions.
- XGBoost provided a strong balance between performance and interpretability, making it a practical alternative for real-world use.
- Traditional Logistic Regression lagged, confirming that linear models struggle with complex, multi-factorial re- lationships.

B. Calibration and Reliability

The calibration curves below in [Fig. 1] to [Fig. 5] show how well each model's predicted probabilities align with actual outcomes:

- The LSTM(Fig. 1) and XGBoost(Fig. 2) models showed strong alignment, with prediction curves closely matching the ideal diagonal.
- The Logistic Regression model showed a tendency to under-predict risk in high-risk cohorts.

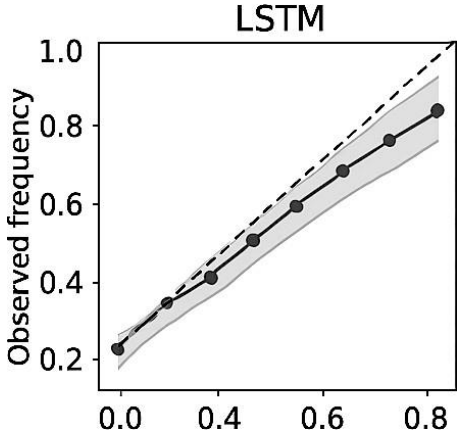

Fig. 1. LSTM Calibration Curve

C. Feature Importance (Interpretability)

SHAP values were used to interpret model decisions and identify influential features.
Top 5 contributors to risk (LSTM + SHAP ensemble anal- ysis):

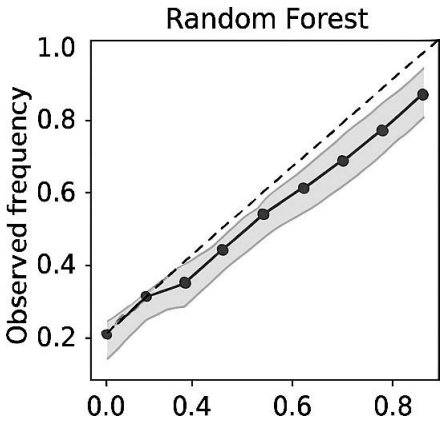
Random Forest
Observed frequency
1.0
0.8
0.6
0.4
0.2
0.0
0.4
0.6
0.8

Fig. 2. XGBoost Calibration Curve

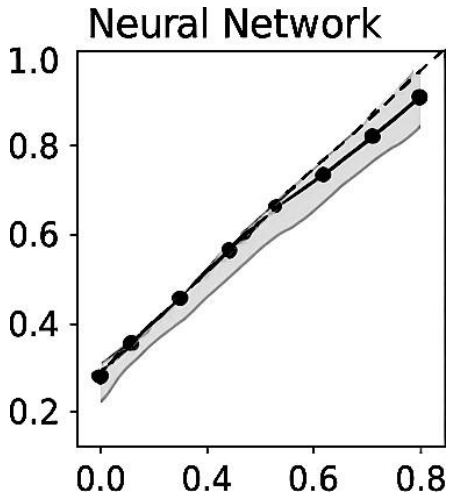

Fig. 3. Neural Network Calibration Curve

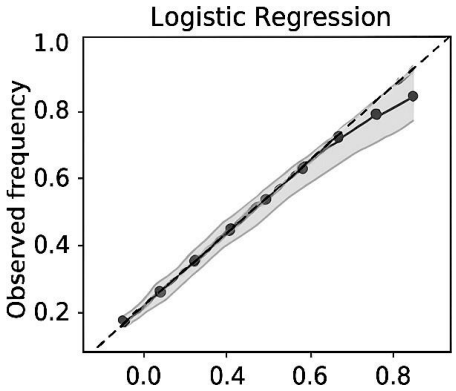

Fig. 4. Logistic Regression Calibration Curve

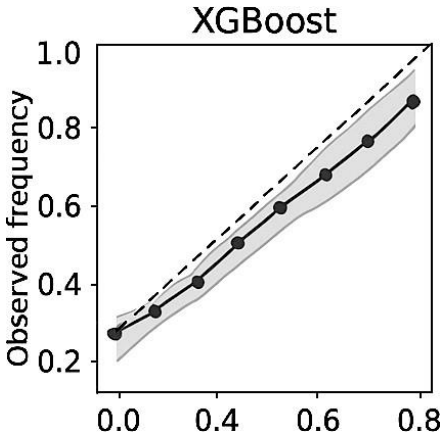

Fig. 5. Random Forest Calibration Curve

1. Number of chronic illnesses
2. Time since last hospitalisation
3. Average claim amount in last 6 months
4. BMI
5. Smoking status

This interpretability framework ensures the system remains compliant with regulatory guidelines around transparency and fairness.

D. Error Analysis

- False Positives: Mostly individuals with high medical expenditure but short-lived health events (e.g., one-time surgery).
- False Negatives: Younger individuals with no prior med- ical history but high future risk due to behavioural or genetic factors.

This insight is vital for understanding model blind spots and guiding hybrid modelling approaches (e.g., mixing rule-based checks with ML predictions).

DISCUSSION

- **Key Findings**
  - Temporal Data Matters: Models that could process se- quential data (like LSTM) significantly outperformed others, confirming the hypothesis that claim patterns over time are strong indicators of future risk.
  - AI Models Generalise Better: The deep learning models achieved higher AUC and F1 scores, demonstrating robustness even on complex, non-linear datasets.
  - Traditional Models Are Limited: Logistic regression, while easy to interpret, failed to match the accuracy or reliability of tree-based or neural models, especially in detecting nuanced patterns.

Practical Implications

- For Insurers: The improved risk prediction can and will lead to better premium pricing, customer segmentation, and reduced frauds.
- For Policymakers: The regulatory bodies need to ensure the fairness, privacy, and explainability of these predictive models
- For Customers: Fairer, more tailored premium structures may be offered, potentially rewarding healthier lifestyles with reduced premiums.

Challenges and Limitations

- Data Quality and Availability: Real-world insurance data may be noisy or incomplete. Synthetic data simulation was used in part, which may not capture all edge cases.
- Bias and Fairness: Historical bias in healthcare access or claims approval might influence predictions. Mitigating algorithmic bias is a future direction.
- Model Interpretability: While tools like SHAP help explain black-box models, there is still resistance in regulated industries to deploy non-transparent systems.

Future Work

- Federated Learning for Privacy: Implementing privacy- preserving ML using federated approaches to train mod- els across institutions without sharing raw data.
- Multimodal Data Fusion: Combining wearable sensor data, genomics, and social determinants of health with claims data for a holistic risk profile.
- Real-time Monitoring: Transitioning from one-time pre- diction to continuous risk scoring using streaming data and edge deployment in mobile health apps.

CONCLUSION

The integration of Artificial Intelligence (AI) into health insurance risk prediction signifies a pivotal transformation in the actuarial domain [9]. Our study demonstrates that AI models—particularly deep learning and sequence- based approaches like Long Short-Term Memory (LSTM) networks—can significantly outperform traditional mod- els in terms of accuracy, flexibility, and predictive insight [5]. These models effectively capture non-linear relation- ships, temporal dependencies, and interactions among features that are typically overlooked by conventional statistical methods.

By augmenting risk assessment with AI, insurers can develop more granular and personalised pricing strategies, improve fraud detection, and enhance claims manage- ment [13]. Furthermore, AI systems can support early interventions by identifying high-risk individuals before claims escalate, enabling preventative care strategies that are mutually beneficial for insurers and policyholders. However, while the performance benefits are substantial, implementation must be approached with caution. Ethical considerations such as data privacy, model bias, and regulatory compliance remain critical [12]. Techniques like differential privacy, federated learning, and inter- pretable AI (e.g., SHAP, LIME) are crucial enablers for responsible adoption. Additionally, collaboration between data scientists, actuaries, legal teams, and healthcare professionals is essential to ensure the developed models are fair, explainable, and clinically relevant.

Future directions for this research include deploying AI models in real-time risk engines, integrating multimodal data from wearable devices and genomics, and exploring reinforcement learning for dynamic policy optimisation. With continuous innovation and responsible stewardship, AI has the potential to redefine the foundations of risk management in health insurance.

REFERENCES

1. A. Rajkomar, E. Oren, K. Chen, et al., “Scalable and accurate deep learning with electronic health records,” npj Digital Medicine, vol. 1, no. 1, pp. 1–10, 2018.
2. J. Friedman, T. Hastie, and R. Tibshirani, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed. Springer, 2009.
3. T. Chen and C. Guestrin, “XGBoost: A Scalable Tree Boosting System,” in Proc. 22nd ACM

SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2016, pp. 785–794.
4. S. M. Lundberg and S.-I. Lee, "A Unified Approach to Inter- preting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, 2017.
5. H. Harutyunyan, H. Khachatrian, D. C. Kale, G. Ver Steeg, and Galstyan, "Multitask Learning and Benchmarking with Clinical Time Series Data," Scientific Data, vol. 6, no. 1, 2019.
6. P. Sendak, M. D'Arcy, J. Kashyap, et al., "A Path for Translation of Machine Learning Products into Healthcare Delivery," npj Digital Medicine, vol. 3, 2020.
7. S. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018.
8. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?" Explaining the Predictions of Any Classifier," in Proc. 22nd ACM SIGKDD, 2016.
9. D. C. Parkes and M. P. Wellman, "Economic Reasoning and Artificial Intelligence," Science, vol. 349, no. 6245, pp. 267–272,2015.
10. A. Esteva, K. Chou, S. Yeung, et al., "Deep Learning-Enabled Medical Computer Vision," npj Digital Medicine, vol. 4, no. 5, 2021.
11. J. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dis- secting Racial Bias in an Algorithm Used to Manage the Health of Populations," Science, vol. 366, no. 6464, pp. 447–453, 2019.
12. T. Mitchell, "The Need for AI Transparency and Explainability in Healthcare," Communications of the ACM, vol. 61, no. 11, pp. 33–36, 2018.
13. M. K. Baig, N. Arshad, and N. Imran, "Machine Learning-Based Risk Assessment in Health Insurance: A Systematic Literature Review," IEEE Access, vol. 10, pp. 12967–12982, 2022.
14. Y. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, "Deep Learning for Healthcare: Review, Opportunities and Challenges," Briefings in Bioinformatics, vol. 19, no. 6, pp. 1236–1246, 2018.
15. S. A. Kharroubi and J. Gabrani, "Health insurance risk prediction using machine learning techniques: A review," in *Proc. 2018 7th Mediterranean Conf. on Embedded Computing (MECO)*, Budva, Montenegro, 2018, pp. 1–4. doi: 10.1109/MECO.2018.8406010.