International Journal of Research in Education: Theory and Practices
ISSN: 3048-8486
Volume 3 Issue 4, July-Aug, 2026

An Intelligent Machine Learning Framework for Health Insurance Claim Cost Prediction Using Healthcare Risk Factors

M.Anitha[1], Patibandla Sushmitha[2]
[1]Assistant Professor, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.
[2]MCA Student, Department of MCA, Megha Institute of Engineering and Technology for Women, Edulabad (Village), Ghatkesar (Mandal), Medchal District, Telangana – 501301, India.

Abstract. For insurance companies to maximize premium computation, enhance financial planning, and facilitate efficient risk management, accurate cost projections of health insurance claims are crucial. The intricate links between lifestyle factors, health issues, and demographic traits that are frequently difficult for conventional statistical prediction tools to capture might occasionally hinder prediction performance. Using a variety of healthcare-related factors, such as age, gender, body mass index (BMI), smoking habits, number of dependents, genetic illnesses, occupation, and residential area, the intelligent machine learning method presented in this study predicts health insurance claim expenses. To increase forecasting accuracy and decision-making, the suggested approach combines feature engineering, predictive modeling, and meticulous data pretreatment. Standard performance metrics like R-squared (R^2) and Mean Absolute Error (MAE) are used to create and assess regression approaches including Linear Regression, Support Vector Regression (SVR), Random Forest Regressor, and Gradient Boosting Regressor. An experimental study found that while Gradient Boosting also produces comparable forecasting performance, the Random Forest Regressor offers the highest prediction accuracy by successfully capturing nonlinear linkages among health-related indicators. Insurance businesses can improve pricing strategies, lower financial uncertainty, anticipate future claim costs more accurately, and construct customized insurance policies more easily with the help of the suggested framework. By offering scalable, dependable, and data-driven prediction capabilities appropriate for contemporary health insurance administration systems, the developed method further illustrates the useful advantages of machine learning for intelligent healthcare analytics.

keywords: Health Insurance, Insurance Claim Prediction, Machine Learning, Healthcare Analytics, Random Forest Regressor, Gradient Boosting Regressor, Support Vector Regression, Linear Regression, Predictive Analytics, Cost Forecasting, Healthcare Risk Assessment, Data Mining.

I. Introduction

The rapid growth of the healthcare industry and the increasing cost of medical treatment have made health insurance an essential component of modern healthcare systems. Health insurance protects individuals from unexpected medical expenses while enabling insurance providers to manage financial risks through appropriate premium planning and claim estimation. As the number of insured individuals continues to increase, insurance companies are required to process a large volume of claims efficiently and accurately. Predicting insurance claim costs has therefore become an important task for insurers because accurate estimation supports premium calculation, financial planning, fraud prevention, and effective risk management [1], [2].

Statistical models and historical claim analysis are the main foundations of traditional insurance cost estimation methods. While these approaches offer a fundamental understanding of claim trends, they frequently fall short of capturing the intricate and nonlinear interactions between lifestyle, medical, and demographic factors. Future healthcare costs are significantly influenced by a number of factors that interact in different ways, including age, gender, body mass index (BMI), smoking habits, genetic anomalies, occupation, residential area, and family medical history. As a result, traditional prediction models often generate erroneous claim projections, which could impact premium rates and raise the financial uncertainty of insurance companies [3], [4].

Innovative methods that can uncover hidden patterns in massive healthcare datasets have been made possible by recent developments in machine learning (ML). In contrast to conventional statistical methods, machine learning algorithms do not require predetermined mathematical assumptions in order to automatically uncover complex correlations between various input elements and insurance claim prices. Artificial Neural Networks (ANNs), Support Vector Regression (SVR), Random Forest Regressor, Gradient Boosting Regressor, XGBoost, and Linear Regression are regression-based learning models that have shown encouraging results in predicting healthcare expenditures with improved accuracy and generalization capability [5], [6]. These sophisticated algorithms can effectively evaluate massive amounts of healthcare data while adapting to a variety of patient features and evolving medical trends.

Because ensemble learning algorithms can increase prediction accuracy by merging many decision models, they have attracted a lot of interest among the current machine learning techniques. Algorithms like Random Forest and Gradient Boosting are excellent for predicting insurance claims because they prevent overfitting, reduce prediction variance, and capture nonlinear relationships. Furthermore, by identifying the primary healthcare risk variables influencing insurance prices, appropriate data preprocessing, feature engineering, categorical data encoding, and correlation analysis greatly enhance model performance [7], [8].

Accurate prediction of insurance claim costs offers several practical benefits for insurance providers, healthcare organizations, and policyholders. Reliable forecasting enables insurers to develop personalized premium structures, improve financial resource allocation, optimize claim settlement procedures, and enhance customer satisfaction through fair pricing strategies. Additionally, intelligent prediction systems assist insurance companies in identifying high-risk individuals, detecting abnormal claim patterns, and supporting strategic business decisions based on data-driven insights [9], [10]. As healthcare datasets continue to grow in size and complexity, machine learning has become an effective solution for improving operational efficiency and reducing uncertainty in insurance management.

Overall, the integration of machine learning into health insurance analytics provides a scalable and reliable approach for predicting medical claim costs with greater precision. Intelligent prediction models support evidence-based decision-making, improve financial planning, strengthen risk assessment, and contribute to the development of transparent and customer-centric insurance services. As healthcare systems continue to evolve, advanced machine learning techniques will play an increasingly important role in enhancing insurance management and promoting efficient healthcare financing [5], [9], [15].

II. Literature Survey

1. S. Suresh, R. Mary Chittilappilly, and S. Shanmugam [2023]

R. Mary Chittilappillya number of predictive algorithms to improve insurance cost estimation and demonstrated how optimization-based learning approaches might increase prediction accuracy while promoting better premium budgeting and financial decision-making.

2. V. Rotta, V. Pagadala, S. R. Bommareddy, V. G. S. K. R. K. Jamili, and S. Vuddanti [2024]

S. Vuddanti et al. introduced a tailored insurance pricing strategy based on machine learning. Their research looked at client health profiles and demographic information to more accurately forecast insurance premiums. The study's conclusions indicate that insurance companies can use advanced prediction models to improve their pricing methods and reduce their financial risk.

3. L. Xiaoqun and L. Run [2022]

L. Xiaoqun and L. Run created a hybrid prediction model that combines Support Vector Machine (SVM) and K-Means clustering to estimate health insurance prices. By merging individuals with similar health characteristics before generating predictions, their approach enhanced predictive skills and discovered significant variables like age, smoking status, and Body Mass Index (BMI).

4.[2022] J. B et al. developed a secure framework for medical insurance cost prediction that protected medical data using machine learning techniques and encryption.

Their experimental review found that Random Forest Regression outperformed conventional regression models in prediction accuracy while appropriately handling sensitive patient data.

5. V. Vijayalakshmi, K. Panimalar, and A. Selvakumar [2023]
V. Vijayalakshmi et al. employed a number of regression approaches, including to forecast medical insurance expenditures. Their comparison study revealed that Random Forest was the most reliable algorithm for estimating healthcare expenses due to its exceptional ability to simulate nonlinear interactions.

6. developed a Cross Gradient Regression Model (CGRM) to predict health insurance costs using lifestyle and demographic data. They used data preparation, normalization, and cross-validation approaches to improve prediction accuracy and reduce forecasting mistakes in healthcare datasets.

7. estimating medical insurance costs. They discovered that smoking habits, age, and BMI were significant predictors of insurance claim costs in their study, which assessed various regression techniques. The study emphasized the potential benefits of machine learning for actuarial analysis and insurance pricing.

8. S. Ravali, G. S., K. P. Linija, D. Mavaluru, and S. A. L. Mary [2023]
rediction. Their work shown that LSVM can handle healthcare datasets with a range of medical and demographic factors by using feature selection and categorical data encoding to improve learning performance.

9. Their results showed that ensemble learning models consistently delivered more reliable and accurate claim cost estimation when compared to traditional regression techniques.

10. to estimate medical insurance fund balances. Their study demonstrated the value of predictive analytics in financial planning and emphasized the importance of accurate forecasting for long-term healthcare insurance management.

III. Research Methodology

The proposed research methodology presents an intelligent machine learning framework for estimating health insurance claim costs by utilizing demographic, medical, and lifestyle-related information. The methodology is organized into several sequentia The primary objective is to develop a reliable predictive model capable of estimating future insurance claim expenses with high accuracy, thereby supporting insurance providers in premium calculation, financial planning, and risk assessment.

The methodology begins with data acquisition, where a health insurance dataset containing multiple customer attributes is collected from a publicly dependents, hereditary

diseases, occupation, residential city, and insurance claim amount. These attributes represent the major health and demographic factors that influence medical expenses and insurance claims. Since real-world healthcare datasets often contain incomplete or inconsistent information, the collected data undergoes detailed preprocessing before model development.

After training, each predictive model generates estimated insurance claim costs for the testing dataset. The predicted values are then compared with the actual claim amounts using standard regression evaluation metrics, includingComparative analysis is subsequently performed to identify the most effective regression algorithm for healthcare insurance cost estimation.

IV. Existing System

Although these models provide a simple mechanism for estimating claim amounts, they often assume linear relationships among variables and are unable to accurately represent the complex interactions between demographic characteristics, lifestyle habits, hereditary conditions, and healthcare expenditures. As a result, prediction accuracy decreases when handling large and diverse healthcare datasets that contain nonlinear patterns.

To improve forecasting performance, several studies have introduced insurance claims and provide better prediction capability than conventional statistical approaches. However, many existing models are developed using a limited number of health-related attributes and are often evaluated on specific datasets, which restricts their ability to generalize effectively across different populations and healthcare environments. In addition, some models require extensive feature engineering and parameter tuning, increasing computational complexity and reducing implementation efficiency.

Another limitation of current insurance claim prediction systems is the insufficient preprocessing of healthcare data categorical variables that require proper transformation before machine learning algorithms can be applied. Several existing approaches do not adequately address these data quality issues, leading to reduced model reliability and increased prediction errors. Furthermore, many prediction frameworks focus on a single machine learning algorithm without performing comprehensive comparisons among multiple regression

Although recent machine learning approaches have significantly improved prediction performance compared with traditional statistical methods, challenges remain in achieving high forecasting accuracy, minimizing prediction errors, and efficiently modeling the nonlinear relationships that exist among healthcare variables. These limitations indicate the need for a more intelligent and comprehensive prediction framework that integrates effective data preprocessing, feature engineering, and comparative evaluation of multiple machine learning algorithms. Such a framework can provide more accurate health insurance claim estimation, support personalized premium calculation,

improve financial planning, and enable data-driven decision-making for modern health insurance management systems.

V. Proposed System

The framework is designed to assist insurance providers in estimating future claim expenses more accurately, thereby supporting premium calculation, financial planning, and risk assessment. Unlike traditional statistical approaches that rely on fixed mathematical assumptions, the proposed system utilizes data-driven learning techniques capable of identifying complex nonlinear relationships among various healthcare attributes. This enables the prediction model to generate more reliable insurance cost estimates while adapting to different customer health profiles.

The prediction process begins with the collection of health insurance data containing important customer information hereditary diseases, occupation, residential city, and insurance claim amount. Since raw healthcare data may contain missing values, duplicate records, inconsistent entries, and categorical variables, comprehensive preprocessing is performed before model training. The preprocessing stage includes data cleaning, missing value imputation, duplicate removal, normalization, and categorical feature encoding. These operations improve data quality and ensure that the learning algorithms receive consistent and meaningful input for accurate prediction. Correlation analysis and feature engineering techniques are further applied to identify the healthcare variables that have the greatest influence on insurance claim costs.

Following preprocessing, establishes a baseline prediction model by estimating linear relationships between input variables and claim costs. Support Vector Regression captures nonlinear patterns using kernel-based learning methods, while Random Forest Regressor employs an ensemble of Regressor sequentially combines several weak prediction models to minimize forecasting errors and improve overall regression performance.

After model training, the developed regression models estimate insurance claim costs for the testing dataset. The predicted claim values are compared with the actual insurance costs using standard evaluation metrics, including among healthcare variables. The Gradient Boosting Regressor also demonstrates competitive performance, while Linear Regression and Support Vector Regression exhibit comparatively lower prediction capability for this dataset.

The proposed framework further supports visual analysis by comparing actual and predicted claim costs through graphical representations. These visualizations allow insurance analysts to evaluate model performance, identify prediction trends, and understand the behavior of different regression algorithms. Such analytical capabilities assist insurance companies in making informed decisions regarding policy pricing, claim management, financial planning, and customer risk profiling.

Overall, the proposed system integrates healthcare with systematic data preprocessing and performance evaluation, the proposed approach delivers more accurate and reliable health insurance claim cost estimation than conventional prediction methods. The framework provides a scalable, efficient, and optimized risk management, enhanced financial decision-making, and better customer service through intelligent healthcare analytics.

VI. System Architecture

Prediction Output. These modules work sequentially to process healthcare information, develop predictive models, and estimate insurance claim costs with high accuracy. The overall framework enables insurance companies to analyze customer health profiles efficiently, improve premium calculation, and support data-driven financial decision-making.

The process begins with the Data Source module, where health insurance information dependents, hereditary diseases, occupation, residential city, and insurance claim amount. These features represent the primary risk factors that influence healthcare expenses and serve as the input variables for the prediction models.

After data collection, the information is forwarded to the Data Preprocessing module, where several operations are performed hereditary diseases, and city are converted into numerical representations using suitable encoding methods. Finally, feature scaling and normalization are applied so that all variables contribute proportionally during model training, improving prediction accuracy and computational efficiency.

The preprocessed dataset is then passed to the Feature Engineering and Selection module. At this stage, input variables and insurance claim costs. Feature importance analysis determines which healthcare attributes have the greatest influence on prediction performance. Only the most relevant variables are selected for model development,

Following feature selection, the dataset enters the Data Splitting module, where itheir ability to estimate insurance claim costs for previously unseen records.

The prepared training dataset is then processed by the Model Development and Model Training modules. Four supervised machine learning regression algorithms are implemented, namely Linear Regression serves as the baseline model for establishing simple relationships between customer attributes and claim costs. Support Vector Regression models nonlinear relationships using kernel functions, whereas Random Forest Regressor cRegressor sequentially refines prediction performance by minimizing the errors generated during previous learning stages. Each model is trained using historical healthcare data to accurately estimate future insurance claim expenses.

After training, the models are assessed in the Model Evaluation module using standard regression performance metrics, including the highest prediction accuracy with the

lowest prediction error, demonstrating its effectiveness in modeling complex healthcare data.

Finally, the selected model generates the Prediction Output, where the estimated insurance claim cost is presented for each customer based on the provided healthcare information. These predictions assist insurance companies in calculating personalized premiums, evaluating customer risk profiles, optimizing financial planning, and improving claim management. The predicted results also support strategic decision-making by enabling insurers to develop fair pricing policies and allocate financial resources more effectively.

Overall, the proposed system architecture integrates healthcare data collection, preprocessing, feature engineering, machine learning model development, performance evaluation, and intelligent claim cost prediction into a unified framework. By combining multiple regression algorithms with systematic data analysis techniques, the architecture provides a reliable, scalable, and efficient solution for modern health insurance claim prediction. The framework enhances prediction accuracy, reduces estimation errors, supports personalized insurance services, and enables data-driven decision-making for healthcare insurance organizations.

VII. Conclusion

The proposed machine learning-based health insurance claim cost prediction framework provides an intelligent and reliable approach for estimating future medical insurance expenses using demographic, medical, and lifestyle-related information. By integrating comprehensive data preprocessing, feature engineering, and multiple regression algorithms, the framework effectively models the complex relationships that exist among healthcare variables influencing insurance claim costs. Unlike conventional statistical approaches, the proposed system utilizes data-driven learning techniques to improve prediction accuracy, enabling insurance organizations to estimate claim expenses more precisely and support evidence-based financial decision-making.

while maintaining the lowest prediction error because of its strong capability to model nonlinear relationships among healthcare attributes. The Gradient Boosting Regressor also produces reliable forecasting results, whereas Linear Regression and Support Vector Regression provide comparatively lower predictive performance for complex healthcare datasets. These findings confirm

The proposed framework also highlights mproving the overall reliability of healthcare analytics. Accurate insurance claim prediction enables insurers to optimize premium calculation, strengthen financial planning, improve customer risk assessment, and support personalized insurance policy development. Furthermore, intelligent prediction systems help insurance providers reduce financial uncertainty, allocate resources efficiently, and enhance operational decision-making through data-driven insights.

Overall,Its ability to generate accurate claim cost estimates supports both insurance companies and policyholders by promoting fair premium determination and effective risk management. Future research may further enhance the prediction framework by incorporating and additional socioeconomic and clinical variables to improve prediction accuracy, transparency, and adaptability for next-generation intelligent health insurance systems.

References

1. "Machine Learning Insights into Personalized Insurance Pricing," S. Vuddanti, V. G. S. K. R. K. Jamili, S. R. Bommareddy, V. Rotta, and V. Pagadala, 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS), Coimbatore, India, 2024, pp. 923–927, doi:10.1109/ICSCSS60660.2024.10625562.
2. "Insurance Prediction of Medical Resource Using Cross Gradient Regression Model," A. Sanjai and K. S. Rekha, IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 934–937, doi:10.1109/IC2PCT60090.2024.10486392.
3. "Prediction of Insurance Cost Through Machine Learning Structured Algorithm," A. Sharma and R. Jeya, IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), Greater Noida, India, 2024, pp. 495–500, doi:10.1109/IC2PCT60090.2024.10486304.
4. Z. Liu, H. Li, and Z. Chen, "Prediction on the Accumulated Balance of Chinese Provincial Employee Medical Insurance Fund Based on GM (1,1) Model," 7th International Symposium on Autonomous Systems (ISAS), Chongqing, China, 2024, pp. 1–5, doi:10.1109/ISAS61044.2024.10552467.
5. M. Rithik and S. Chokkalingam, "Image-Based Classification for Vehicle Accident Detection and Insurance Claim Using Artificial Intelligence," International Conference on Emerging Systems and Intelligent Computing (ESIC), Bhubaneswar, India, 2024, pp. 521–525, doi:10.1109/ESIC60604.2024.10481545.
6. "Utilizing a New Approach in a Multiple Linear Regression Model to Predict Insurance Charges," H. Khajeh, M. Keshavarz, and M. Ehsani, 2024 20th CSI International Symposium on Artificial Intelligence and Signal Processing (AISP), Babol, Iran, 2024, pp. 1–6, doi:10.1109/AISP61396.2024.10475251.
7. R. Mary Chittilappilly, S. Suresh, and S. Shanmugam, "A Comparative Analysis of Optimizing Medical Insurance Prediction Using Genetic Algorithm and Other Machine Learning Algorithms," International Conference on Advances in Computing, Communication, and Applied Informatics (ACCAI), Chennai, India, 2023, pp. 1–6, doi:10.1109/ACCAI58221.2023.10199979.
8. "Implementation of Medical Insurance Price Prediction System Using Regression Algorithms," V. Vijayalakshmi, A. Selvakumar, and K. Panimalar, 5th International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2023, pp. 1529–1534, doi:10.1109/ICSSIT55814.2023.10060926.
9. "Medical Insurance Cost Analysis and Prediction Using Machine Learning," International Conference on Innovative Data Communication Technologies and Applications (ICIDCA), Uttarakhand, India, 2023, pp. 113–117,

doi:10.1109/ICIDCA56705.2023.10100057, T. Thejeshwar, S. H. T., V. K. V., and K. R.

10. N. H. Vardhan, R. M., V. R. S. V. Matcha, S. R. S. Reddy, and V. N. Chowdary, "UHI Web Application: Empowering Healthcare with Disease Prediction for Seamless Doctor Consultations," IEEE Global Conference for Advancement in Technology (GCAT), Bengaluru, India, 2023, pp. 1–5, doi:10.1109/GCAT59970.2023.10353384.
11. L. Xiaoqun and L. Run, "An Improved K-Means Clustering Model Based on Support Vector Machine for Health Insurance Cost Prediction," IEEE 2nd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2022, pp. 521–525, doi:10.1109/ICETCI55101.2022.9832085.
12. "End-to-End Encryption and Prediction of Medical Insurance Cost," J. B., A. Ghosh, and J. A. Kumar R., 6th International Conference on Trends in Electronics and Informatics (ICOEI), Tirunelveli, India, 2022, pp. 846–850, doi:10.1109/ICOEI53556.2022.9777238.
13. L. Qin and B. Liu, "Agricultural Insurance Demand Forecasting Model Based on Support Vector Machine," 14th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA), Changsha, China, 2022, pp. 725–728, doi:10.1109/ICMTMA54903.2022.00149.
14. C. Jyothsna, K. Srinivas, B. Bhargavi, A. E. Sravanth, A. T. Kumar, and J. N. V. R. S. Kumar, "Health Insurance Premium Prediction Using XGBoost Regressor," International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2022, pp. 1645–1652, doi:10.1109/ICAAIC53929.2022.9793258.
15. "Explainable Artificial Intelligence for Healthcare Insurance Risk Prediction," A. Kumar, P. Sharma, and R. Singh, IEEE Access, vol. 13, pp. 28451–28467, 2025.
16. Y. Chen, H. Zhao, and X. Wang, "Ensemble Machine Learning Models for Medical Cost Prediction Using Electronic Health Records," Expert Systems with Applications, vol. 257, Art. no. 124812, 2025.
17. "Artificial Intelligence-Based Healthcare Cost Prediction: A Comprehensive Survey," Computers in Biology and Medicine, vol. 180, Art. no. 109214, 2025, M. Alqahtani, S. Rahman, and A. Khan.
18. "Deep Learning Framework for Healthcare Expenditure Prediction Using Clinical and Demographic Data," J. Lee, K. Kim, and S. Park, IEEE Access, vol. 13, pp. 45126–45142, 2025.
19. "Machine Learning Techniques for Intelligent Health Insurance Claim Prediction and Risk Assessment," Journal of Healthcare Informatics Research, vol. 9, no. 2, pp. 145–164, 2025; R. Gupta, N. Patel, and S. Mehta.