Journal of Global Research in Multidisciplinary Studies (JGRMS)
Volume 2 | Issue 10 | October 2026 | Pages 09–15
DOI: https://doi.org/xxxxxxxxxxxxx
e-ISSN: 3049-2963
RESEARCH ARTICLE OPEN ACCESS
Available online: https://saanvipublications.com/journals/index.php/jgrms

Machine Learning-Based Insurance Claim Prediction Using Demographic and Health Factors

Ms. Shivani Jain
Assistant Professor
Department of Computer Science and Applications
Mandsaur
shivani.jain@meu.edu.in

Corresponding Author: * Shivani Jain (shivani.jain@meu.edu.in)

Received: 25 September, 2026 | **Revised:** 3 October 2026 | **Accepted:** 08 October 2026

***Abstract*—In today's healthcare environment, the increasing cost of medical treatment, medicines, and hospitalization has created significant financial challenges for individuals and families. Health insurance offers financial security by covering qualified medical expenditures and facilitating access to critical healthcare services. However, the growing volume of insurance claims and variations in demographic and health conditions require efficient and accurate approaches for Claim Prediction and decision-making. This study uses the Insurance Claim Analysis: Demographic and Health Dataset, which has 1,340 records and 11 characteristics, to offer a machine learning-based method for forecasting health insurance claims. The data were preprocessed through missing-value handling, duplicate removal, noise reduction, label encoding, SMOTE-based balancing, Min–Max normalization, and stratified data splitting. The proposed Light Gradient Boosting Machine (Light GBM) model was trained using demographic and health-related features and evaluated using accuracy, precision, recall, and F1-score. The proposed model achieved 98.8% accuracy, 98.9% precision, 98.6% recall, and 98.7% F1-score, demonstrating its effectiveness for insurance claim prediction. The outcomes indicate that the proposed approach can support accurate and efficient analysis of health insurance claims and facilitate data-driven decision-making in insurance applications.**

***Keywords*—Health Insurance, Insurance Claim Prediction, Machine Learning, Predictive Analytics, Cost Forecasting, Healthcare Risk Assessment, Data Mining.**

I. INTRODUCTION

Healthcare is a fundamental component of human well-being, yet access to affordable medical treatment remains a challenge due to rising healthcare expenses. Health insurance protects individuals from unexpected financial burdens by covering hospitalization, medications, diagnostic procedures, and medical treatments [1][2]. However, policyholders may face complex policy conditions, variations in hospital charges, lengthy claim procedures, processing delays, and uncertainty regarding claim amounts. Therefore, accurate insurance claim prediction is essential for improving transparency, financial planning, and risk management [3][4][5][6].

The increasing number of insured individuals and healthcare claims has made accurate claim estimation an important requirement for insurance companies [7][8]. Traditional approaches rely mainly on statistical techniques and historical claim analysis, but insurance claim amounts are influenced by multiple interacting factors. Age, gender, BMI, Smoking Habits, number of dependents, occupation, Residential Area, and medical conditions can significantly affect healthcare expenditure. Considering demographic and health-related factors together can therefore provide a more comprehensive representation of insurance risk and expected claim amounts [9][10].

Accurate claim prediction can benefit both policyholders and insurance providers by supporting financial forecasting, resource allocation, risk assessment, premium planning, and efficient claim management. It can also help identify unusual claim patterns and reduce potential financial losses. However, the complex relationships among demographic characteristics, Health Conditions, lifestyle factors, and medical expenses make conventional approaches less effective in capturing nonlinear patterns [11][12][13][14]. Hence, data-driven approaches are required to learn meaningful relationships from historical insurance and healthcare data [15][16].

ML and DL have emerged as powerful data-driven approaches for addressing the complexities of insurance claim prediction by learning hidden patterns and nonlinear relationships from large healthcare datasets [17][18]. ML techniques can analyze demographic, medical, lifestyle, and historical claim information to estimate claim amounts, assess risk, identify important influencing factors, and improve prediction accuracy [19]. DL further enhances predictive capability by automatically learning complex feature interactions and high-dimensional relationships from healthcare data [20][21]. Therefore, the proposed Machine Learning-Based Insurance Claim Prediction Using Demographic and Health Factors focuses on utilizing demographic and health attributes to develop an effective predictive framework for accurate insurance claim estimation and risk assessment.

Copyright © 2026 The Author(s).
This is an open access article distributed under the terms of the Creative Commons Attribution 4.0 International (CC BY 4.0) License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, adaptation, and reproduction in any medium, provided the original author(s) and the source are properly credited.

DOI: https://doi.org/xxxxxxxxxxxx

A. Motivation and Contribution

The motivation for this study is the growing need for accurate and efficient insurance claim prediction, driven by rising healthcare costs and claim complexity. The analysis of demographic and health-related factors is important as they strongly affect healthcare costs and the outcome of insurance claims. Traditional approaches may struggle to capture the complicated interactions that these elements might have. As such, an ML approach is investigated to enhance reliable claim prediction, risk assessment and effective insurance administration. This research offers several key contributions as listed below:

- Utilizes demographic and health-related factors for insurance claim prediction.
- Performs data preprocessing, including cleaning, encoding, normalization, and duplicate removal.
- Applies SMOTE to address Class Imbalance and improve representation of minority claim categories.
- Develops a Light Gradient Boosting Machine (Light GBM) model for insurance claim classification.
- Evaluates the model using accuracy, precision, recall, F1-score, and ROC-AUC.
- Provides a data-driven approach to support efficient and accurate insurance claim analysis.

B. Justification and Novelty

Accurate insurance claim prediction is essential for supporting efficient and consistent decision-making in insurance services. Demographic and health-related attributes are used in the study to uncover meaningful patterns related to claim outcomes. To handle class imbalance, SMOTE is used, and to capture complex relationships among the input features, Light GBM is used. The novelty is the combination of systematic preprocessing, class balancing, and Light GBM in a single framework for reliable insurance claim prediction.

C. Paper Organization

The remainder of this paper is structured as follows: Section II discusses existing studies related to insurance claim prediction based on demographic and health factors. Section III describes the dataset, preprocessing techniques, and proposed Light GBM model implementation. Section IV presents the experimental findings and comparative performance analysis. Finally, Section V summarizes the major outcomes of the study and outlines potential directions for future research.

II. Literature Review

The literature review examines the growing use of ML in insurance claim classification, cost prediction, feature analysis, and automated decision support, leveraging diverse health-related factors.

Rao et al. (2026) present a unified machine learning framework for insurance claim classification and claim amount prediction using integrated health and motor insurance data. The approach employs Random Forest, XGBoost, and a stacking classifier with preprocessing and feature standardization. To address class imbalance, SMOTE-equivalent manual oversampling is applied only to training data. Without oversampling, the model achieves 93% accuracy but limited claim detection recall. After oversampling, minority claim recall improves while maintaining competitive accuracy of 91% [22].

Mohan et al. (2025) aim to estimate health insurance claim costs using ML based on Health Risk factors, including age, sex, BMI, smoking status, and city. A Kaggle health insurance dataset is used to benchmark different algorithms through data preprocessing, feature engineering, model selection, R^2, and Mean Absolute Error. The outcomes showed that RF performed the best among the evaluated models, with 96.7% accuracy, owing to its capacity to represent complicated interactions [23].

Ramani et al. (2024) developed an appropriate ML model for forecasting insurance claim amounts and calculating individuals' out-of-pocket expenses. The study evaluates RFR, MLR, XGBoost Regression, GBR, and DTR for health insurance claim prediction. Experimental results demonstrated accurate predictions, with Random Forest achieving 96.7% accuracy compared with the other models [24].

Sharma and Jeya (2024) examine factors including age, sex, BMI, smoking, and number of children to predict health insurance costs. Linear regression is used to analyze the relationship between these factors and insurance costs. The dataset was partitioned with 70% for training, achieving 81.3% accuracy. The study provides insights into post-COVID-19 health insurance spending and supports healthcare cost forecasting and policy decision-making [25].

K et al. (2023) propose a machine learning solution for health insurance claims by identifying key data sources, examining healthcare data properties, and analyzing feature patterns through univariate and bivariate analysis. Data visualization is used to identify impactful features. After preprocessing, four algorithms—LR, RF, DTC, and Naive Bayes—are compared. The Decision Tree Classifier achieved the highest accuracy of 97.03% [26].

Vijayalakshmi, Selvakumar and Panimalar, (2023) used a dataset with 24 features, including all relevant attributes needed for prediction of insurance cost. The implementation done using regression algorithms such as Linear Regression, DTR, Lasso Regression, Ridge Regression, RFR, Elastic Net Regression, Support Vector Regression, K Nearest Neighbor Regression and Neural Network Regression in R Programming. There were seven metrics applied to measure the performance of the model, namely MSE, RMSE, MAE, Mean Absolute Percentage Error (MAPE), R-Squared (R2), Adjusted R-Squared (Adj. R2), and Explained Variance Score (EVS). Random Forest Regression outperformed with 0.9533 as RS quired value. This prediction system will reduce manual work and give the prediction accurately in medical field [27].

Recent studies demonstrate that machine learning can effectively support insurance claim classification and cost prediction using demographic, health, and behavioral factors. RF, DT, XGBoost, and regression-based models have achieved strong predictive performance across different datasets. However, existing research generally focuses on either claim classification or cost estimation separately, limiting integrated analysis of insurance-related outcomes. Several studies also provide limited consideration of class imbalance, feature interactions, model interpretability, and robust evaluation across diverse claimant characteristics. These shortcomings highlight the need for an all-encompassing prediction framework that enhances accurate and understandable insurance claim prediction by including pertinent demographic and health variables. Table I

DOI: https://doi.org/xxxxxxxxxxxx

summarizes the reviewed studies, datasets, algorithms, evaluation approaches, and reported findings

TABLE I. RECENT RESEARCH ON HEALTH AND DEMOGRAPHIC FACTORS IN INSURANCE CLAIM PREDICTION

Author	Proposed Work	Dataset	Key Findings	Limitations & Future Work
Rao et al. (2026)	Develops a unified ML framework for insurance claim classification and claim-amount prediction using Random Forest, XGBoost, and stacking.	Integrated health and motor insurance data	Accuracy was 93% without oversampling and 91% after oversampling, with improved minority-class recall.	Combines health and motor insurance and emphasizes class imbalance; demographic and health-specific predictors are not the central focus.
Mohan et al. (2025)	Predicts health insurance claim costs using ML based on age, sex, BMI, smoking status, and city.	Kaggle health insurance dataset	Random Forest achieved 96.7% accuracy, demonstrating the usefulness of demographic and health-risk factors for insurance-cost prediction.	Uses a relatively conventional dataset and focuses mainly on cost forecasting. Future work can investigate additional health and demographic variables and advanced models.
Ramani et al. (2024)	Predicts health insurance claim amounts using Random Forest Regression, Multiple Linear Regression, XGBoost Regression, Gradient Boosting, and Decision Tree Regression.	Health insurance claim dataset	Random Forest achieved approximately 96.7% accuracy and showed strong capability for claim-amount prediction.	Primarily addresses claim amount rather than claim occurrence/classification and may not fully exploit demographic-health interactions.
Sharma & Jeya (2024)	Uses Linear Regression to predict health insurance costs from age, sex, BMI, smoking status, and number of children.	Health insurance dataset	Reported 81.3% accuracy using a 70% training split and highlighted relationships between demographic/health variables and insurance costs.	Linear regression may not capture complex nonlinear relationships; future research can evaluate ensemble and deep-learning methods.
K et al. (2023)	Applies Logistic Regression, Random Forest, Decision Tree, and Naive Bayes for health insurance claim analysis and classification.	Health insurance claims dataset	Decision Tree achieved the highest reported accuracy of 97.03%.	Limited model diversity and comparatively limited exploration of advanced feature engineering and explainability.
Vijayalakshmi, Selvakumar & Panimalar (2023)	Predicts insurance costs using multiple regression and ML models, including Linear Regression, Random Forest, SVR, KNN, and Neural Network Regression.	Insurance-cost dataset with 24 features	Random Forest Regression achieved an R^2 of 0.9533, outperforming the other regression models.	Focuses on cost prediction rather than claim classification; future work could integrate richer demographic and health information.

Research Gaps: Existing studies have mainly focused on insurance, claim-cost estimation, or general claim classification using different ML techniques. Limited research has explored the combined influence of demographic and health-related factors for multi-class insurance claim prediction. The effects of class imbalance and systematic preprocessing on claim classification have also received comparatively less attention. Therefore, an integrated approach using data balancing and Light GBM is required to improve insurance claim prediction and risk analysis.

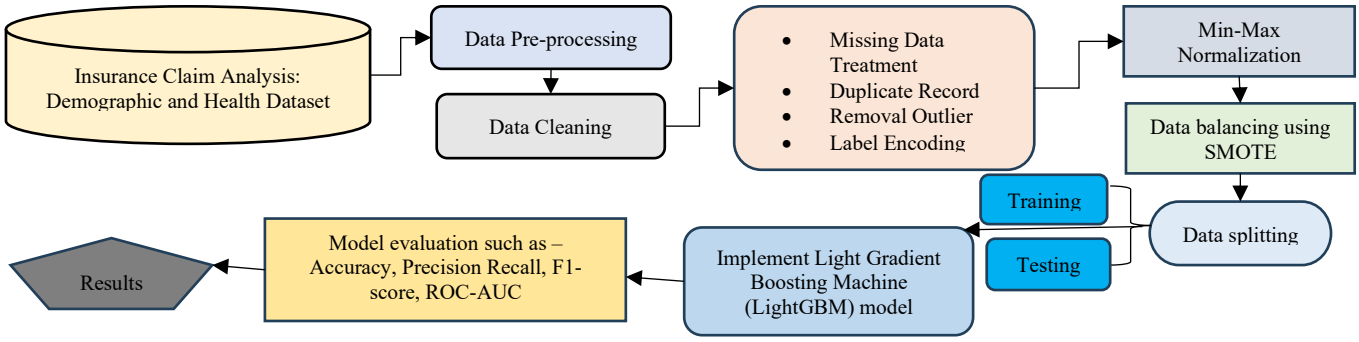


Fig. 1. Proposed Workflow for Machine Learning-Based Insurance Claim Prediction Using Demographic Factors

III. RESEARCH METHODOLOGY

The proposed methodology starts with data cleaning, label encoding, class balancing using SMOTE, Min–Max normalization, and an 80–20 stratified split. The preprocessed Insurance Claim Analysis: Demographic and Health Dataset is used to train the Light Gradient Boosting Machine (Light GBM) model for insurance claim prediction based on demographic and health-related attributes. The model performance is evaluated using accuracy, precision, recall, F1score, Confusion Matrix, and ROC-AUC to assess its classification effectiveness. Fig.1 shows the proposed workflow for predicting insurance claims from demography and health aspects using machine learning techniques.

The following section describes each stage of the proposed methodology in detail:

A. Data Collection and Exploration

The study utilizes the Insurance Claim Analysis: Demographic and Health Dataset, comprising 1,340 records and 11 attributes, to analyze insurance claims based on demographic and health-related factors. The dataset includes relevant personal, demographic, and health information for predicting insurance claim outcomes using machine learning techniques. Data visualizations such as bar plots and heatmaps were used to examine claim distribution and feature correlations, as shown below:

J. Glob. Res. Multidiscip. Studies, 2 (10):09-15
DOI: https://doi.org/xxxxxxxxxxxx

Fig. 2 shows the distribution of three types of claims: "Low", "Medium", and "High". The vast majority of the records are in the "Low" category, which overwhelmingly dominates with a number of almost 1200. The “Medium” category is drastically lower, at about 150 instances. Meanwhile, the "High" category is nearly non-existent with a negligible count close to zero, highlighting a strongly skewed dataset.

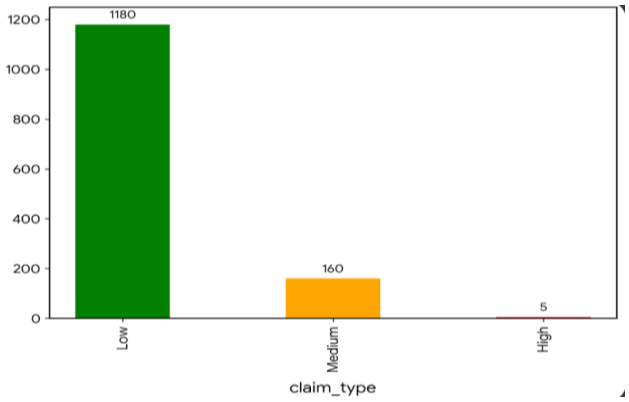


Fig. 2. Bar Graph of Class Distribution

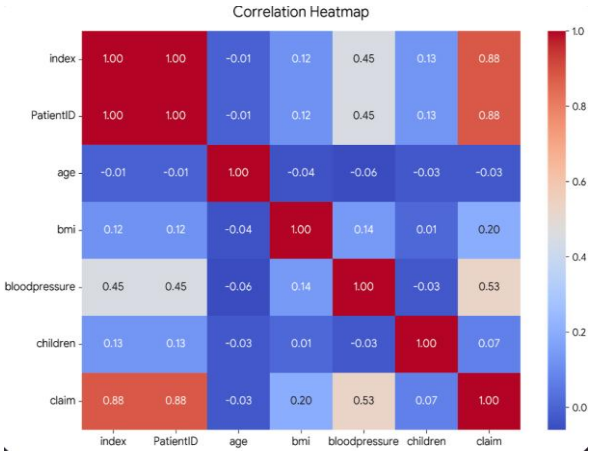


Fig. 3. Correlation Matrix Heatmap

Fig. 3 illustrates the linear relationships among numerical attributes in the dataset, with values ranging from -0.06 to 1.00. There is a strong positive correlation (0.88) between claim and both index and Patient ID and a moderate correlation (0.53) between claim and blood pressure. On the other hand, there are no significant non-zero correlations between the variables for age and children. Overall, the main features linked to the claim values are blood pressure, index, and Patient ID.

B. Data Pre-Processing

The Insurance Claim Analysis: Demographic and Health dataset was prepared through a systematic process of data integration, cleaning, and feature transformation. The preprocessing stage involved treating missing values, removing duplicate and irrelevant records, eliminating outliers, applying label encoding to categorical attributes, and normalizing numerical features. The key preprocessing steps adopted in this study are outlined below:

C. Data Cleaning

Data cleaning was performed by identifying and removing missing values, incomplete records, and duplicate entries to maintain dataset consistency and reliability. Outliers were detected and removed to minimize the influence of abnormal observations on the analysis. These cleaning steps improved data quality and supported more accurate and reliable ML model performance.

D. Label Encoding

All categorical variables were converted into numerical representations using label encoding, making them compatible with ML algorithms. This transformation enabled the model to process categorical attributes efficiently while preserving their class-wise distinctions.

E. Min-Max Normalization

The numerical features were normalized using the Min–Max scaling method, which transforms the values to a range of 0 to 1. This scaling helps maintain a consistent feature range and improves the efficiency of the classification process. The normalization is expressed by the following mathematical Equation (1):

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \quad (1)$$

where X represents the original value of the feature, X' is the normalized value, X_{min} is the minimum value of the feature, and X_{max} is the maximum value of the same

F. Data Balancing Using SMOTE

Data balancing is an important preprocessing step used to reduce class imbalance and ensure that each class is adequately represented during model training. In this study, SMOTE was applied to generate synthetic instances for the minority class. SMOTE identifies the nearest neighboring minority samples and creates new instances by interpolating between them, thereby increasing minority-class representation and producing a more balanced dataset for reliable model training.

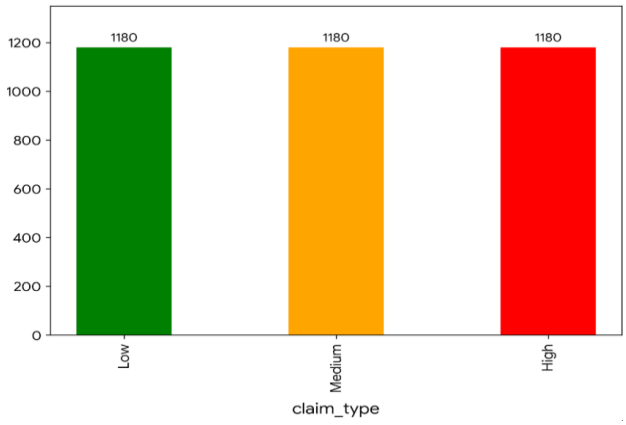


Fig. 4. Bar Graph of Class Distribution After SMOTE

Fig. 4 illustrates the balanced distribution of the claim type variable following the application of SMOTE oversampling. The frequency of all three categories "Low", "Medium" and "High" has been equalized by synthesizing artificial samples for the minority classes, with 1,180 instances in each category. This change eliminates the initial class imbalance in the dataset and ensures that each class has a similar data structure, improving the model's learning performance.

G. Data Splitting

The dataset was split into training and testing sets using an 80:20 stratified ratio, preserving the original class distribution in both subsets.

DOI: https://doi.org/xxxxxxxxxxxxx

H. Proposed Light Gradient Boosting Machine (Lightgbm) Model

This study proposes a supervised ML approach based on the Light Gradient Boosting Machine (Light GBM) for forecasting insurance claims using demographic and health-related factors. Light GBM is a gradient boosting framework that builds an ensemble of DT sequentially, where each new tree focuses on reducing the errors made by the previous trees. The prediction of the Light GBM model is obtained by combining the outputs of multiple weak learners, represented by Equation (2):

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \qquad f_k \in F \tag{2}$$

where $\hat{y}_i$ represents the predicted output, K is the number of trees, x_i is the input feature vector, and f_k denotes the kth decision tree. The model parameters are optimized by minimizing an objective function consisting of the training loss and a regularization term, as expressed in Equation (3):

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{3}$$

where $l(y_i, \hat{y}_i)$ is the loss between the actual and predicted values, while $\Omega(f_k)$ controls model complexity to reduce overfitting. LightGBM employs an efficient leaf-wise tree growth strategy and histogram-based learning, which enables fast training and effective handling of large and high-dimensional datasets. The final prediction is obtained by aggregating the learned tree outputs, making LightGBM suitable for intelligent recruitment classification and hiring-decision prediction tasks. The proposed LightGBM model was configured with 200 estimators, a learning rate of 0.05, and a maximum tree depth of 10 to provide effective classification. The model used 31 leaves, a minimum of 20 samples per leaf, and feature fraction of 0.8 for controlled tree growth and improved generalization. Additionally, bagging fraction was set to 0.8, bagging frequency to 5, and random state to 42 to ensure stable and reproducible training.

I. Evaluation Metrics

The performance of the proposed model was evaluated using several standard classification metrics. A confusion matrix was generated to present the classification outcomes by indicating the numbers of correctly and incorrectly predicted instances for each class. The matrix was utilized to determine the True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). These values were subsequently used to calculate accuracy, precision, recall, and F1-score, as presented below:

Accuracy: The fraction of occurrences that were properly categorized relative to all instances. It is a measure of how well the model predicts class labels as a whole. It is given as Equation (4)

$$Accuracy = \frac{TP+TN}{TP+Fp+TN+FN} \tag{4}$$

Precision: The proportion of true positives (instances that were successfully predicted) to the total of false positives (instances that were mistakenly forecasted). It measures the accuracy of optimistic forecasts. How good the classifier is in predicting the positive classes is expressed as Equation (5)

$$Precision = \frac{TP}{TP+FP} \tag{5}$$

Recall: The proportion of true positives to the total of false negatives (missed cases). It shows that the model can identify every positive case. In mathematical form it is given as Equation (6)

$$Recall = \frac{TP}{TP+FN} \tag{6}$$

F1 score: The harmonic mean of perc and rec provides a balanced estimate of both measurements. It is a single statistic that describes the model's performance. Mathematically, it is given as Equation (7)

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{7}$$

Receiver Operating Characteristic Curve (ROC): The ROC curve graphically illustrates the model's classification performance across different decision thresholds by plotting the True Positive Rate (TPR) against the False Positive Rate (FPR). TPR, also known as sensitivity or recall, represents the proportion of positive instances correctly identified, whereas FPR denotes the proportion of negative instances incorrectly classified as positive and is calculated as 1 − specificity.

IV. Results and Discussion

This section outlines the experimental configuration and assesses the suggested model's performance, emphasizing its efficacy and computing efficiency throughout the tests under consideration.

A. Experimental Setup

The proposed model was developed and tested on a high-performance computing system featuring an Intel Core i9-13900K processor at 3.0 GHz, 64 GB DDR5 RAM, and an NVIDIA RTX 4090 GPU with 24 GB VRAM. The experimental environment used Windows 11 Pro and Python-based libraries, including Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn, to perform data preprocessing, visualization, model implementation, and performance evaluation.

B. Experimental Results

The proposed model was trained on the Insurance Claim Analysis: Demographic and Health dataset and tested using the following classification metrics: accuracy, precision, recall, and F1-score as shown in Table II. The Light Gradient Boosting Machine (Light GBM) achieved an accuracy of 98.8%, indicating a high level of classification accuracy. It obtained 98.9% precision and 98.6% recall, indicating reliable identification of the respective classes with minimal classification errors. Moreover, the model achieved an F1 score of 98.7%, showing a good balance between precision and recall for insurance claim prediction.

TABLE II. Classification Results of Proposed Model Demographic and Health-Based Insurance Claim Prediction

Matrix	Proposed Light Gradient Boosting Machine (Light GBM)
Accuracy	98.8
Precision	98.9
Recall	98.6
F1-score	98.7

Fig. 5 confusion matrix of the proposed Light GBM model for three classes (0, 1, and 2). The model correctly classified 236 samples of Class 1 and 27 samples of Class 2, indicating strong classification performance. Only a small number of misclassifications were observed, with 4 Class 2 samples predicted as Class 1 and 1 Class 0 sample predicted as Class 2. Overall, the matrix demonstrates that Light GBM effectively distinguishes the three classes with limited classification errors.

J. Glob. Res. Multidiscip. Studies, 2 (10):09-15
DOI: https://doi.org/xxxxxxxxxxxx

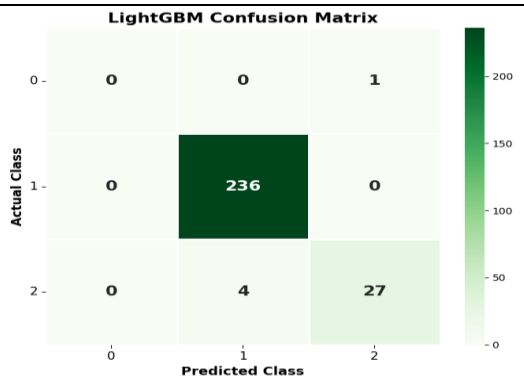


Fig. 5. Confusion Matrix for the Lightgbm Model

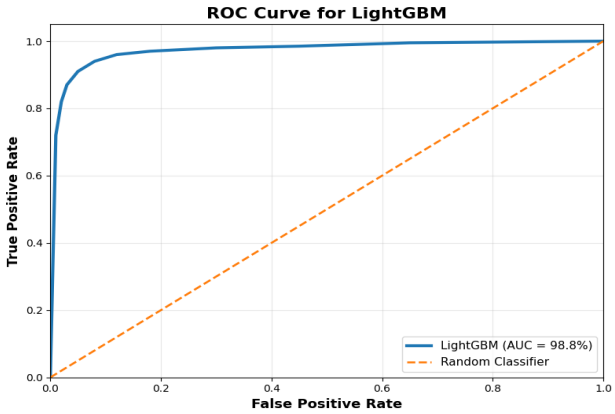


Fig. 6. ROC Curve for Light GBM Model

Fig. 6 illustrates the ROC curve of the Light GBM model for insurance claim prediction. The model achieves an AUC of 98.8%, indicating strong class-discrimination capability. The ROC curve remains close to the upper-left region, reflecting a high true-positive rate with a low false-positive rate. Overall, the results demonstrate the model's effective ability to distinguish between the target classes.

C. *Comparative Analysis*

To assess the effectiveness of the proposed Light Gradient Boosting Machine (Light GBM) model, its classification performance was compared with existing machine learning models, as presented in Table III. Logistic Regression (LR) achieved 82% accuracy, 81% precision, 88% recall, and 85% F1-score, while XGBoost obtained 83%, 73%, 83%, and 75%, respectively. K-Nearest Neighbors (KNN) demonstrated improved performance with 92.99% accuracy, 93.5% precision, 92.99% recall, and 92.97% F1-score. In comparison, the proposed Light GBM achieved an accuracy of 98.8%, precision of 98.9%, recall of 98.6%, and F1-score of 98.7% on the insurance claims classification task using demographic and health factors, outperforming the other models.

TABLE III. COMPARATIVE ANALYSIS OF MACHINE LEARNING MODELS FOR INSURANCE CLAIM PREDICTION USING DEMOGRAPHIC AND HEALTH FACTORS

Model	Accuracy	Precision	Recall	F1-score
LR [28]	82	81	88	85
XGBoost [29]	83	73	83	75
KNN [30]	92.99	93.5	92.99	92.97
Proposed LightGBM	98.8	98.9	98.6	98.7

D. *Performance Discussion*

The proposed Light Gradient Boosting Machine (Light GBM) achieved an accuracy of 98.8% for insurance claim prediction. This result indicates that the model can correctly classify a high proportion of insurance claim instances based on demographic and health factors. Compared with the other evaluated machine learning models, Light GBM demonstrates improved classification accuracy. Thus, the proposed model provides an effective approach for accurate insurance claim prediction.

V. CONCLUSION AND FUTURE STUDY

Insurance is a financial protection mechanism that helps individuals manage potential losses related to health, vehicles, property, education, business, and other unforeseen circumstances. Of these, one of the most crucial is health insurance as rising healthcare costs and health risks can be a heavy burden. In this research, machine learning was used to forecast insurance claim results based on demographic and health-related factors. The proposed Light Gradient Boosting Machine (Light GBM) achieved an accuracy of 98.8%, outperforming LR (82%), XGBoost (83%), and KNN (92.99%). These results demonstrate that Light GBM can effectively capture relationships among demographic and health attributes and provide accurate insurance claim classification. The proposed method hence provides a data-driven framework to aid efficient and reliable insurance claim prediction.

A. *Limitation and Future Work*

The study has limitations due to the size and scope of the demographic and health-related data available, which could limit the generalizability of the findings to other insurance groups. Future work can incorporate larger real-world datasets, additional health and socioeconomic factors, and advanced machine learning or deep learning techniques to f a

- **Funding:** This research received no external funding.
- **Institutional Review Board Statement:** Not applicable.
- **Informed Consent Statement:** Not applicable.
- **Data Availability Statement:** The data supporting the findings of this study are available from the corresponding author upon reasonable request.
- **Conflicts of Interest:** The author declares no conflict of interest.

How to cite this article:
S. Jain, "Machine Learning-Based Insurance Claim Prediction Using Demographic and Health Factors," *Journal of Global Research in Multidisciplinary Studies*, vol. 2, no. 10, pp. 09–15, Oct. 2026, Doi: XXXX.

REFERENCES

[1] S. Chowdhury, P. Mayilvahanan, and R. Govindaraj, "Optimal feature extraction and classification-oriented medical insurance prediction model: machine learning integrated with the internet of things," *Int. J. Comput. Appl.*, vol. 44, no. 3, pp. 278–290, Mar. 2022, doi: 10.1080/1206212X.2020.1733307.

[2] C. S. Pareek, "Testing Approaches in Life Insurance: Accelerated and Fluidless Underwriting Amidst Data-Driven Dynamics," *Int. J. Sci. Res. Publ.*, vol. 14, no. 11, pp. 8–10, Nov. 2024, doi: 10.29322/IJSRP.14.11.2024.p15503.

[3] A. Alnuaimi, A. Alshehhi, K. Salah, R. Jayaraman, I. A. Omar, and A. Battah, "Blockchain-Based Processing of Health Insurance Claims for Prescription Drugs," *IEEE Access*, vol. 10, pp. 118093–118107, 2022, doi: 10.1109/ACCESS.2022.3219837.

[4] U. Orji and E. Ukwandu, "Machine learning for an explainable

cost prediction of medical insurance," *Mach. Learn. with Appl.*, vol. 15, no. 21, p. 100516, Mar. 2024, doi: 10.1016/j.mlwa.2023.100516.

[5] D. P. Dāsepalli, "Artificial Intelligence in Digital Banking: A Systematic Literature Review and a Novel Cross-Domain Resilience Framework for Customer Experience, Fraud Detection, and Risk Management," Jul. 16, 2026. doi: 10.21203/rs.3.rs-10362812/v1.

[6] V. P. Rambabu, J. Sreerama, and J. T. S. Singh, "AI-Driven Data Integration: Enhancing Risk Assessment in the Insurance Industry," *Aust. J. Mach. Learn. Res. Appl.*, vol. 2, no. 2, pp. 130–179, July, 2022.

[7] N. Dhieb, H. Ghazzai, H. Besbes, and Y. Massoud, "A Secure AI-Driven Architecture for Automated Insurance Systems: Fraud Detection and Risk Measurement," *IEEE Access*, vol. 8, pp. 58546–58558, 2020, doi: 10.1109/ACCESS.2020.2983300.

[8] W. B. Zulfikar, K. Aprila, A. F. Huda, N. Lukman, and A. R. Atmadja, "C4.5 and ID3 Comparison to Classify Credit Issue on Indonesian National Health Insurance," in *2022 8th International Conference on Wireless and Telematics (ICWT)*, IEEE, Jul. 2022, pp. 1–4. doi: 10.1109/ICWT55831.2022.9935501.

[9] K. Dutta, S. Chandra, M. K. Gourisaria, and H. GM, "A Data Mining-based Target Regression-Oriented Approach to Modeling of Health Insurance Claims," in *2021 5th International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, Apr. 2021, pp. 1168–1175. doi: 10.1109/ICCMC51019.2021.9418038.

[10] I. Fursov *et al.*, "Sequence Embeddings Help Detect Insurance Fraud," *IEEE Access*, vol. 10, pp. 32060–32074, 2022, doi: 10.1109/ACCESS.2022.3149480.

[11] R. Y. Gupta, S. S. Mudigonda, and P. K. Baruah, "A Comparative Study of Using Various Machine Learning and Deep Learning-Based Fraud Detection Models For Universal Health Coverage Schemes," *Int. J. Eng. Trends Technol.*, vol. 69, no. 3, pp. 96–102, Mar. 2021, doi: 10.14445/22315381/IJETT-V69I3P216.

[12] M. E. Haque and M. E. Tozal, "Identifying Health Insurance Claim Frauds Using Mixture of Clinical Concepts," *IEEE Trans. Serv. Comput.*, vol. 15, no. 4, pp. 2356–2367, Jul. 2022, doi: 10.1109/TSC.2021.3051165.

[13] K. Kapadiya *et al.*, "Blockchain and AI-Empowered Healthcare Insurance Fraud Detection: an Analysis, Architecture, and Future Prospects," *IEEE Access*, vol. 10, pp. 79606–79627, 2022, doi: 10.1109/ACCESS.2022.3194569.

[14] S. Mukherjee, "Machine Learning-Enabled Classification of Diabetes Using Clinical via Lifestyle Health Data," in *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)*, Houston, TX, USA: IEEE, Feb. 2026, pp. 1–6. doi: 10.1109/ICAIC67076.2026.11395709.

[15] K. Kaushik, A. Bhardwaj, A. D. Dwivedi, and R. Singh, "Machine Learning-Based Regression Framework to Predict Health Insurance Premiums.," *Int. J. Environ. Res. Public Health*, vol. 19, no. 13, p. 7898, Jun. 2022, doi: 10.3390/ijerph19137898.

[16] E. Nabrawi and A. Alanazi, "Fraud Detection in Healthcare Insurance Claims Using Machine Learning," *Risks*, vol. 11, no. 9, p. 160, Sep. 2023, doi: 10.3390/risks11090160.

[17] S. Vuddanti, V. G. S. K. R. K. Jamili, S. R. Bommareddy, V. Rotta, and V. Pagadala, "Machine Learning Insights into Personalized Insurance Pricing," in *2024 2nd International Conference on Sustainable Computing and Smart Systems (ICSCSS)*, IEEE, Jul. 2024, pp. 923–927. doi: 10.1109/ICSCSS60660.2024.10625562.

[18] N. K. S. Patel, "Enterprise AI Operations in Healthcare IT Infrastructure," *J. Inf. Syst. Eng. Manag.*, vol. 11, no. 3, pp. 2508–2522, Sep, 2026, doi: https://doi.org/10.5281/zenodo.22773821.

[19] S. Amirineni, "Advanced Healthcare Analytics using AI and ML for Insurance Pattern Detection," in *2026 7th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, Jul. 2026, pp. 2022–2029. doi: 10.1109/ICSSIT69151.2026.11656547.

[20] 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," in *2024 7th International Symposium on Autonomous Systems (ISAS)*, IEEE, May 2024, pp. 1–5. doi: 10.1109/ISAS61044.2024.10552467.

[21] M. Rithik and S. Chokkalingam, "Image-Based Classification for Vehicle Accident Detection and Insurance Claim Using Artificial Intelligence," in *2024 International Conference on Emerging Systems and Intelligent Computing (ESIC)*, IEEE, Feb. 2024, pp. 521–525. doi: 10.1109/ESIC60604.2024.10481545.

[22] M. G. Rao, P. R. H, P. H, A. S. B. M, V. B. S, and S. Shanbhag, "A Unified Explainable Hybrid ML Framework for Insurance Risk Assessment Using Multisource Data Fusion," in *2026 International Conference on Emerging Research in Smart Electronics and Machine Informatics (ECMI)*, IEEE, Apr. 2026, pp. 1–10. doi: 10.1109/ECMI68341.2026.11603025.

[23] S. Mohan, S. Sharma, S. Agrawal, and S. Kamatchi, "Optimization of Insurance Claim Cost Prediction Through Health Data and Machine Learning," in *2025 Fifth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, IEEE, Jan. 2025, pp. 1–7. doi: 10.1109/ICAECT63952.2025.10958928.

[24] K. Ramani, S. T. Kumar, P. P. S. Datta, P. Jamuna, and K. S. Nithin, "Predicting Health Insurance Claim Amount through Machine Learning Algorithms," in *2024 IEEE International Conference on Information Technology, Electronics and Intelligent Communication Systems (ICITEICS)*, IEEE, Jun. 2024, pp. 1–6. doi: 10.1109/ICITEICS61368.2024.10625132.

[25] A. Sharma and R. Jeya, "Prediction of Insurance Cost through ML Structured Algorithm," in *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, IEEE, Feb. 2024, pp. 495–500. doi: 10.1109/IC2PCT60090.2024.10486304.

[26] V. K *et al.*, "Predicting health insurance claim frauds using supervised machine learning technique," in *2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM)*, IEEE, Apr. 2023, pp. 1–7. doi: 10.1109/ICONSTEM56934.2023.10142604.

[27] V. Vijayalakshmi, A. Selvakumar, and K. Panimalar, "Implementation of Medical Insurance Price Prediction System using Regression Algorithms," in *2023 5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, IEEE, Jan. 2023, pp. 1529–1534. doi: 10.1109/ICSSIT55814.2023.10060926.

[28] V. P. Thakre, R. D. Poul, and A. D. Sawarkar, "Predictive Precision: Unraveling Health Insurance Claim Patterns With Logistic Regression and Decision Trees," *Cureus J. Comput. Sci.*, Mar. 2025, doi: 10.7759/s44389-025-03010-y.

[29] K. Mavundla, S. Thakur, E. Adetiba, and A. Abayomi, "Predicting Cross-Selling Health Insurance Products Using Machine-Learning Techniques," *J. Comput. Inf. Syst.*, vol. 66, no. 5, pp. 525–542, Sep. 2026, doi: 10.1080/08874417.2024.2395913.

[30] D. B. Olawade, A. Osborne, A. A. Soladoye, O. E. Oluwadare, E. O. Awogbindin, and O. Z. Wada, "Smart insurance analytics: A novel ensemble feature selection approach to unlock health insurance coverage predictions in Sierra Leone.," *Int. J. Med. Inform.*, vol. 211, no. January, p. 106313, May 2026, doi: 10.1016/j.ijmedinf.2026.106313.