

Research Article

Deep Learning-Based Medical Insurance Cost Prediction Using Personal and Health Attributes

Mr. Ram Pratap Singh

Department of Computer Science and Engineering, Lakshmi Narain College of Technology, Bhopal



Received: 19 September 2026
Revised: 28 September 2026
Accepted: 14 October 2026
Published: 26 October 2026

Doi: 10.55640/ijdsml-06-02-13
Page No: 229-240

Copyright: © 2026 Authors retain the copyright of their manuscripts, and all Open Access articles are disseminated under the terms of the Creative Commons Attribution License 4.0 (CC-BY), which licenses unrestricted use, distribution, and reproduction in any medium, provided that the original work is appropriately cited.

Abstract

The purpose of an insurance policy is to lessen or eliminate the monetary impact of certain risks and unanticipated events. The significance of accurate cost prediction has grown due to the ongoing increase in healthcare prices and the fact that health insurance premiums are affected by several demographic and health-related factors. Insurance companies, lawmakers, and people may all benefit from accurate cost projections for healthcare when making coverage and budgeting decisions. To estimate insurance costs and analyses complicated correlations among health characteristics, machine learning (ML) and deep learning (DL) methods provide useful ways. The purpose of this research was to create ML and DL models that could predict health insurance premiums based on important demographic and health-related characteristics. With $R^2=92.94\%$, $RMSE=0.27$, $MSE=0.0727$, and $MAE=0.143$, the suggested Recurrent Neural Network (RNN) performed admirably. These outcomes prove that the suggested RNN can successfully identify patterns and generate reliable estimates of insurance premiums. Healthcare cost management and data-driven insurance planning can both benefit from the created method.

Keywords: Medical Insurance, Feature Importance, Underwriting, Premium Prediction, Machine Learning, Predictive Modelling, Data-Driven Decision-Making.

I. INTRODUCTION

Insurance is a strategy that helps mitigate or eliminate the monetary burden linked to different kinds of risks and unanticipated events. The significance of accurate cost prediction has grown due to the ongoing increase in healthcare prices and the fact that health insurance premiums are affected by several demographic and health-related factors. Decisions about coverage and financial planning may be better made by consumers, lawmakers, and insurance providers when healthcare expenses are predicted [1][2]. To estimate insurance costs and analyses complicated correlations among health characteristics, machine learning (ML) and deep learning (DL) methods provide useful ways. The purpose of this research was to create ML and DL models that could predict health insurance premiums based on important demographic and health-related characteristics. With $R^2=92.94\%$, $RMSE=0.27$, $MSE=0.0727$, and $MAE=0.143$, the suggested Recurrent Neural Network (RNN) performed admirably. These outcomes prove that the suggested RNN can successfully identify patterns and generate reliable estimates of insurance premiums. Healthcare cost management and data-driven insurance planning can both benefit from the created method [3][4][5].

Historical claim data, statistical methodologies, actuarial analysis, and manual evaluation procedures have traditionally been used to estimate insurance rates. Age, sex, BMI, smoking status, family size, and location are a few of the demographic, health, and lifestyle factors that impact healthcare spending [6][7]. Accurate premium estimation might be problematic due to the complicated interactions among these elements. As a result, medical cost estimation and the discovery of meaningful links within insurance data are two areas where predictive techniques are showing growing promise [8][9].

The process of estimating future insurance costs for medical coverage requires looking at previous data on people and their healthcare-related traits [10]. Individuals can have a better grasp of the elements impacting healthcare cost [11], and insurance firms can use accurate prediction to aid in risk appraisal [12], premium estimation, and policy development. Statistical analysis and data visualization can help uncover important variables, distributions, correlations, and trends in insurance datasets. A better understanding of the key factors that affect health insurance premiums may be gained via this type of study [13][14][15].

Recent advances in ML and DL have brought forth state-of-the-art methods for predictive analysis in insurance and healthcare applications [16][17]. Learned complicated correlations of demographic, health, lifestyle, and personal factors may be discovered using ML and DL systems trained on historical data. Large datasets and better healthcare spending prediction are two areas where these methods shine [18][19]. Medical insurance premiums are predicted in this study using ML and DL techniques [20]. Attributes such as age, BMI, smoking status, sex, number of children, and area are used. Important elements and predictive performance for insurance cost prediction are also assessed in the study.

Motivation and Contribution

The rising expense of healthcare and the subsequent necessity for reliable methods of estimating health insurance premiums according to specific demographic and health-related variables prompted this research. Insurance companies can enhance their risk assessment and people may better understand and budget for their healthcare costs with the aid of accurate premium forecast. Consequently, in order to create a data-driven method for predicting health insurance premiums, this work investigates ML and DL methodologies. The following are a few of the most important contributions made by this study:

- Constructed a model for predicting health insurance premiums based on RNNs and health-related and demographic data.
- Used data preparation techniques, such as addressing missing values, handling outliers, encoding categories, and normalizing using StandardScaler.
- Developed a model for predicting insurance premiums using analysis of variance with an F-test.
- Used R^2 , RMSE, MSE, and MAE measures to assess the suggested model.
- Proven that deep learning can be used to forecast health insurance premiums.
- Offered a data-driven strategy to aid in insurance cost planning and measurement.

Justification and Novelty

The growing need for reliable health insurance cost prediction using demographic and health-related variables is the driving force behind this work. Coming up with an RNN-based model that can understand intricate correlations among many elements impacting insurance rates is where the innovation is. The approach integrates a single prediction framework with data preparation, categorical encoding, feature selection, normalization, and deep learning. When it comes to health insurance premium estimates and insurance planning, this method is dependable and data-driven.

Paper Organization

The rest of the article is organized like this: A review of the literature on health insurance premium prediction is presented in Section II, and the dataset, preprocessing methods, and suggested model implementation are detailed in Section III, Materials and Methodology. The findings and comparisons from the experiments are detailed in Section IV, Results and Discussion. Section V, Conclusion and Future Scope, concludes the study by reviewing its key points and outlining its possible future research areas.

Literature Review

A comprehensive literature evaluation was conducted to identify the current methods, results, and areas where

research on health-based insurance premium prediction is lacking.

Mehta et al. (2026) The goal is to address this difficulty by combining risk-sensitive regression models and interaction-based feature engineering methodologies. A regular model and an improved model including a body mass index (BMI) and smoking interaction (i.e., BMI multiplied by smoking) are compared in the authors' analysis of medical insurance costs using linear regression models. The models are evaluated with 1,338 records of real-world data obtained from the Kaggle Medical Insurance Cost dataset. With a 34% decrease in Mean Absolute Error (MAE) from 4181 to 2757 and a 10% increase in the coefficient of determination (R^2) from 0.783 to 0.865, high-cost cohorts show a considerable improvement in predictive ability [21].

Mishra et al. (2026) intend to integrate interaction-based feature engineering techniques with risk-sensitive regression models in an effort to solve this issue. The authors compare a baseline model that does not account for the interaction between smoking and body mass index (BMI) with an improved model that does so. The authors then use this upgraded model to analyse the costs of health insurance. There are 1,338 records in the Kaggle Medical Insurance Cost data set, which is used to test the models. The coefficient of determination (R^2) increased by 10% from 0.783 to 0.865 and the mean absolute error (MAE) decreased by 34% from 4181 to 2757 in high-cost cohorts, leading to a notable gain in predictive power [22].

Mohan et al. (2025) the goal is to use machine learning algorithms to predict the expenses of health insurance claims according to health risk indicators such as city, age, sex, body mass index (BMI), smoking status, and the like. The accuracy of predictions is evaluated using R^2 and Mean Absolute Error after data preprocessing, feature engineering, and model selection on a Kaggle health insurance dataset. The capacity of Random Forest to capture complicated associations in health data was demonstrated by its best results among the evaluated models, which were 96.7% accuracy [23].

R and Arunraj (2025) suggest a metaheuristic optimization method to enhance the prediction of health insurance premiums by means of hyperparameter adjustment and feature selection. Bayesian Optimization, Genetic Algorithm (GA), and Particle Swarm Optimization (PSO) are evaluated with respect to MAE, MSE, RMSE, and R^3 . Using Random Forest and Gradient Boosting, the suggested approach outperformed the competition, increasing accuracy by 4.3% with an RMSE of 2510.23 and a R^2 of 92.5%. Ensemble models perform better than regression-based methods, with smoking status and body mass index (BMI) being the most important variables to include [24].

Kandula et al. (2024). Healthcare institutions and others with a stake in the industry must have reliable cost forecasting tools at their disposal. Despite the abundance of patient data generated by the healthcare business, very little is done with it to comprehend healthcare expenses. In order to evaluate healthcare insurance expenses and increase policy efficiency, this study utilizes approaches from computational intelligence and machine learning. Due to the difficulty in making a single model selection, the prediction accuracy of many regression models is assessed. A maximum accuracy of 88.98% was attained with Stochastic Gradient Boosting [25].

M. Chittilappilly, et al.(2023) evaluate several machine learning approaches to forecasting healthcare costs using demographic and health-related data. On a preprocessed dataset, TPOT-based automated machine learning, Random Forest, Gradient Boosting, LassoLarsCV, and Linear Regression are tested. The model that was developed by TPOT and which consisted of Gradient Boosting Regressor and LassoLarsCV had the best performance, with a test accuracy of 87.45% and an RMSE of 0.0686 [26].

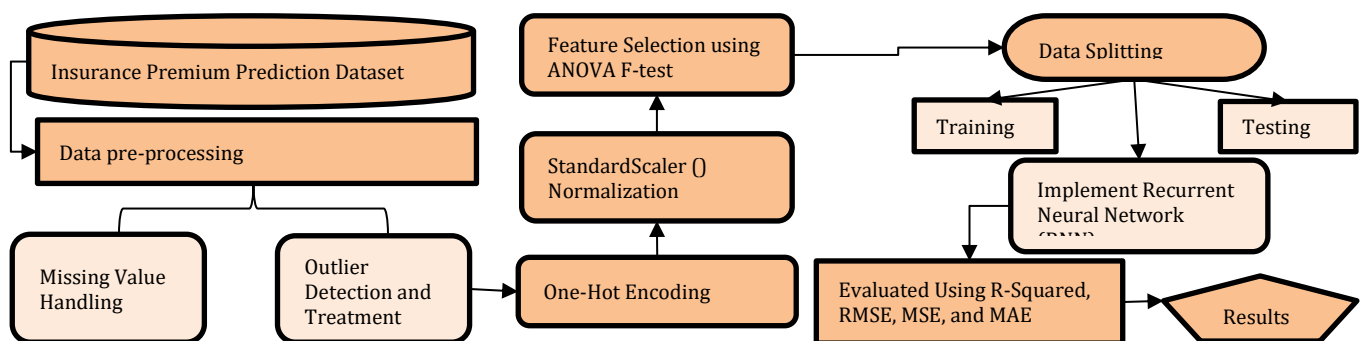
Bhatia et al. (2022) propose a method for predicting health insurance premiums using machine learning. Since the COVID-19 pandemic, health insurance has risen to the forefront of study, and several efforts have been undertaken to address this issue. They drew on a Kaggle dataset with 1338 elements representing US healthcare costs. Age, gender, body mass index (BMI), smoking history, family size, etc. are some of the dataset features utilized to forecast insurance premiums. They also found the correlation between the price and these characteristics by employing linear regression. A 70-30 split was used to train the system, and an accuracy of 81.3% was attained [27].

The methods used, datasets utilized, important results, limitations, and future research directions of recent works on health-based insurance premium prediction are summarized in Table I.

TABLE I. Recent Studies on Insurance Premium Prediction Using Health Attributes

Author	Proposed Work	Dataset	Key Findings	Limitations & Future Work
Mehra et al. (2026)	Proposed an insurance-cost prediction approach using risk-sensitive regression and interaction-based feature engineering, particularly considering the relationship between BMI and smoking.	Kaggle Medical Insurance Cost Dataset; 1,338 records.	Improved R^2 from 0.783 to 0.865 and reduced MAE from 4181 to 2757 for high-cost groups.	Focuses mainly on regression and engineered interactions; broader ML/DL approaches and additional health attributes could be explored.
Mishra et al. (2026)	Developed an ML-based approach for insurance-premium prediction using demographic, health, lifestyle, and environmental factors.	1,338 individuals, including age, BMI, smoking, location, air and water quality.	Reported $R^2 = 0.9654$ for the highest-performing approach; smoking and BMI were identified as important factors.	Dataset is relatively small; further investigation using larger datasets and DL approaches could improve generalization.
Mohan et al. (2025)	Applied ML for predicting health insurance claim costs using personal and health-risk attributes.	Kaggle health-insurance dataset containing age, sex, BMI, smoking status, and city.	Reported 96.7% predictive accuracy for the best-performing approach and highlighted ML for insurance cost forecasting.	Limited attribute diversity and scope; additional health variables and DL-based prediction could be investigated.
R and Arunraj (2025)	Proposed metaheuristic optimization for feature selection and hyperparameter tuning in insurance-cost prediction.	Health insurance cost data involving factors such as BMI and smoking.	Optimized approach achieved R^2 of 92.5% and RMSE of 2510.23, with improved predictive performance.	Emphasis is on optimization rather than DL; larger and more diverse datasets could be considered.
Kandula et al. (2024)	Developed a computational-intelligence approach for estimating healthcare insurance expenditures using ML-based predictive analysis.	Healthcare/insurance expenditure data.	The reported stochastic gradient boosting approach achieved 88.98% accuracy.	Multiple regression approaches and broader datasets could be evaluated; DL-based approaches remain a potential extension.
M. Chittilappilly, et al. (2023)	Compared different ML approaches for predicting medical insurance charges and investigated automated optimization.	Preprocessed medical-insurance dataset containing demographic and health-related information.	The automated approach achieved RMSE = 0.0686 and 87.45% accuracy on the test set.	Focuses mainly on automated ML; DL and richer health-related attributes could be explored.
Bhatia et al. (2022)	Developed an ML-based health-insurance prediction system for estimating insurance costs from personal characteristics.	USA Medical Cost Personal Dataset; 1,338 records.	Used demographic and health-related attributes and reported 81.3% accuracy with a 70:30 split.	Uses a relatively small dataset and limited predictive methodology; DL-based approaches were not investigated.

Research Gaps: There has been scant investigation into deep learning methods for predicting health insurance premiums, with most previous research concentrating on more conventional methods like ensemble neural networks and classical regression. Further limiting the ability to capture complicated interactions among factors impacting the outcome is the fact that many research employ a restricted number of health-related and demographic information. In contrast to the premium prediction, integrated preprocessing, feature selection, normalization, and recurrent learning have received very little attention. Improving health insurance premium prediction necessitates an RNN-based system capable of successfully learning complicated feature correlations.

**Fig. 1. Proposed Framework for Health-Based Insurance Premium Prediction**

Research Methodology

Data collection from the Insurance Premium Prediction Dataset, data preprocessing, handling of missing values, outlier detection, categorical encoding, StandardScaler normalization, feature selection using the ANOVA F-test, and 80:20 data splitting into training and testing sets comprise the suggested methodology. A dense layer, two 100-unit recurrent layers with 30-time steps each, and a final RNN model were trained to forecast health insurance premiums based on demographic and health-related variables. We measured the accuracy and prediction error of the suggested insurance premium prediction model using R^2 , MSE, RMSE, and MAE. The suggested model for predicting health insurance premiums is illustrated in Fig 1. Each component of the suggested technique is described in depth in the section that follows:

Data Gathering and Analysis

The Kaggle Insurance Premium Prediction Dataset was utilized in this work. It contains more than 1,300 records with seven attributes, such as charges, smoking, area, children, BMI, sex, and age. Using the dataset, health insurance premiums were predicted according to demographic and health-related variables. The feature distributions and relationships among the variables were examined using data visualizations, such as bar graphs and heatmaps:

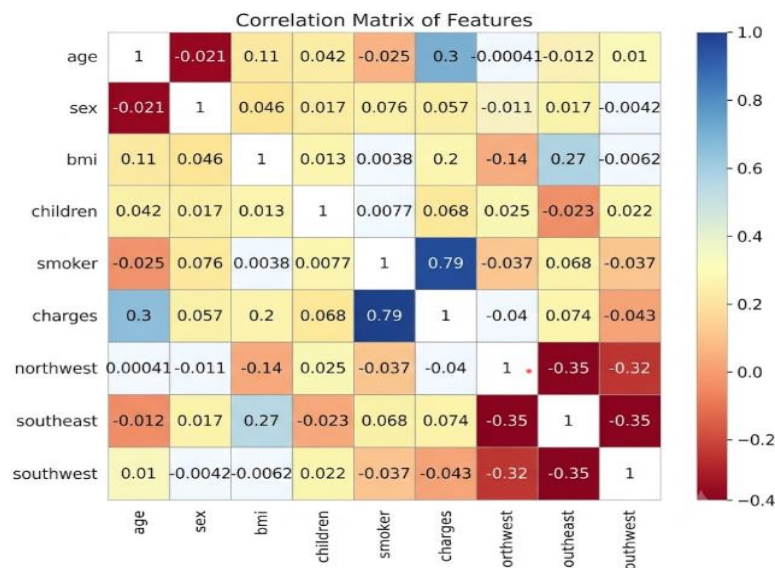


Fig 2. Correlation Matrix Heatmap

There is a robust positive connection of 0.79 between smoker and charges, one of the nine demographic, geographical, and financial factors shown in Fig 2's pairwise correlations. The other geographical categories (northwest, southeast, and southwest) exhibit moderate negative correlations with each other, ranging from -0.32 to -0.35, as they are mutually exclusive. When it comes to other features like age, sex, and kid count, the linear dependency is negligible, with values close to zero.

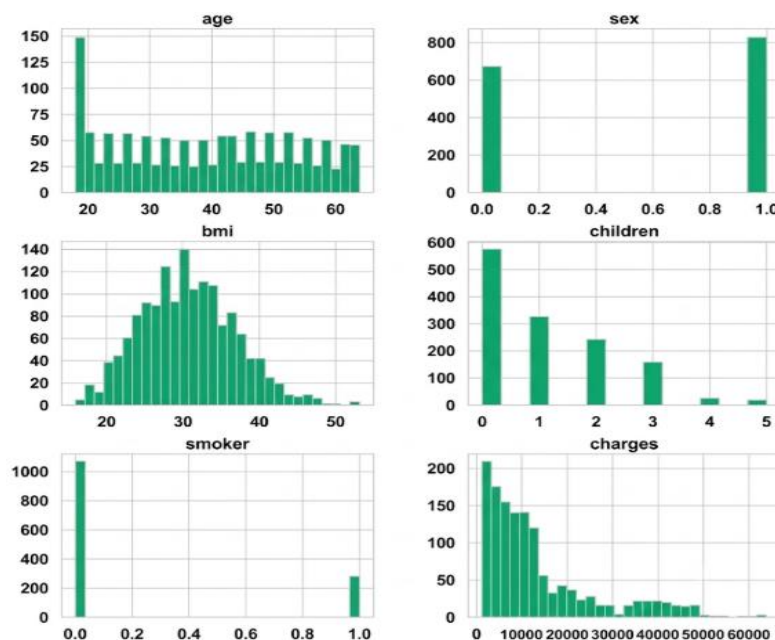


Fig.3. Feature Distribution Histograms

Fig 3 displays the individual probability distributions for all significant demographic and health-related variables. For adults, the distribution is almost even when it comes to age, and for sex and smoking status, it's binary. Medical costs and out-of-pocket expenses for children are right-skewed, with most values being low, but continuous measures show a regularly distributed body mass index (BMI) centered around 30. The subplots provide a clear visual representation of the feature spreads and class balances that are essential for validating baseline data and conducting predictive

modelling.

The distributions and interactions among age, sex, body mass index (BMI), children, smokers, charges, and geographical features are shown in Fig 4. Feature patterns, relationships, and category-specific differences in insurance costs are highlighted.

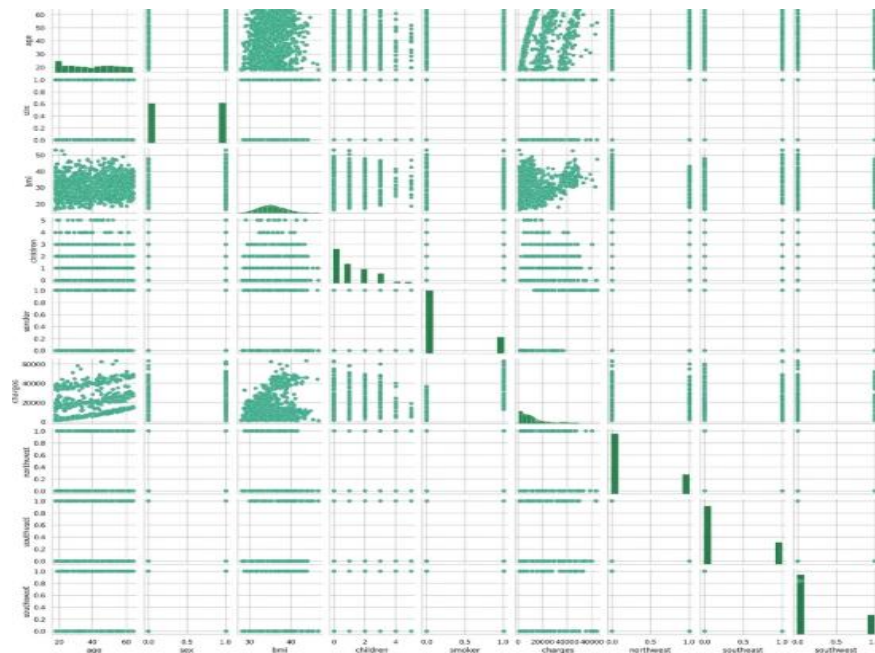


Fig. 4. Feature Pairplot

Data Pre-Processing

The Insurance Premium Prediction Dataset was preprocessed to enhance the quality of the data by addressing inconsistencies, outliers, duplicate records, and missing values. In order to make the suggested machine learning models more efficient and to guarantee consistent scales, the numerical features were normalized. The dataset was then processed and cleaned up to get it ready for training the models and evaluating their performance. What follows is a synopsis of the primary preprocessing steps:

- **Missing Value Handling:** Mean or median imputation was used to handle missing numerical values, while the mode was used to replace missing categorical data. The original data distribution was not substantially affected, and completeness was guaranteed.
- **Outlier Detection and Treatment:** Data parameters including body mass index (BMI), age, and insurance premiums were analyzed using the interquartile range (IQR) approach to spot any outliers. In order to keep relevant data intact and limit their impact, extreme values were limited to reasonable ranges.

Categorical Feature Encoding

The use of One-Hot Encoding allowed for the numerical representation of previously categorized data, including gender, smoking status, and geography. This made it possible for ML systems to efficiently handle categorical data.

StandardScaler () Normalization

The dataset was transformed using the StandardScaler () function to ensure that the resultant distribution had a mean of zero and a standard deviation of one, taking into consideration the varying scales of each descriptor. In order to accomplish this change, we may use Equation (1) to get the average value of each observation and then divide it by the standard deviation:

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

where z represents the feature's modified value, x stands for each descriptor's original value, μ denotes the feature's

mean, and σ denotes its standard deviation in the dataset.

Feature Selection Using ANOVA F-Test

The most important factors that go into predicting insurance premiums were chosen through feature selection. Quantitative characteristics were evaluated for statistical significance and relevance to the dependent variable using the analysis of variance F-test. Fig 5 shows the association coefficients between medical charges, demographic information, and health-related variables in a horizontal bar chart format. The biggest positive link with charges is smoking status, followed by age and BMI; rather modest correlations are observed with gender and geographical factors.

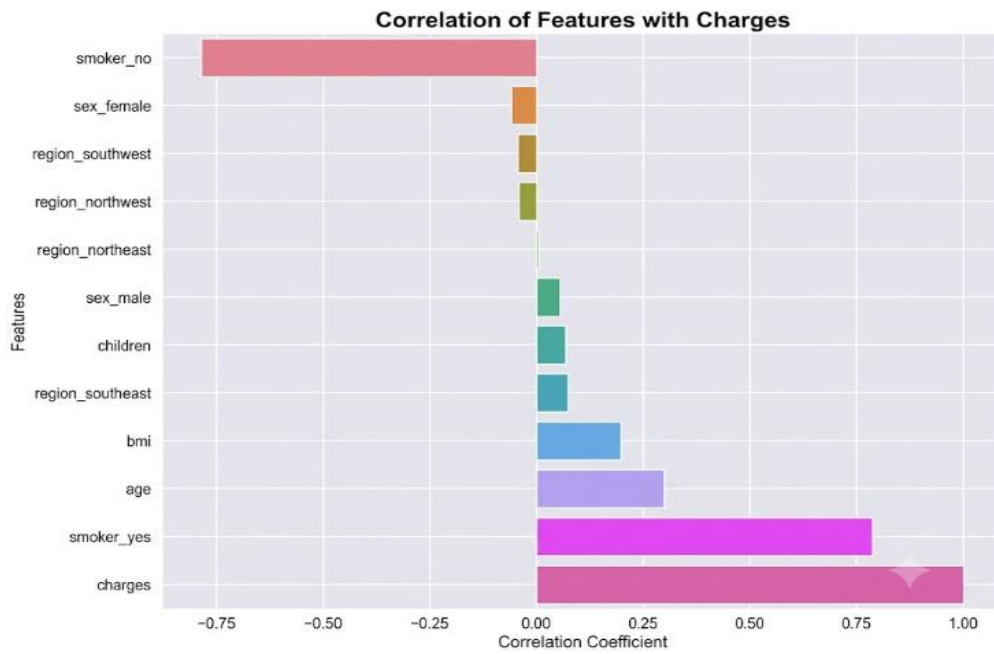


Fig. 5. Correlation of Features With Charges

Data Splitting

There were 80% training data points and 20% test data points in the preprocessed dataset. The model was trained on the training set, and its performance on unseen data was tested on the testing set.

Proposed Recurrent Neural Network (Rnn)

RNN models are able to handle sequential data with ease due to the fact that RNN units are built to remember past states. When calculating the current state, this memory characteristic was taken into consideration. As a result, RNN units take in two values: the past output and the present input value. The RNN units' underlying mechanics are best described as Equation (2).

$$y(t) = f(x(t) + y(t - 1)) \quad (2)$$

When the input is represented by $x(t)$, the output is denoted by $y(t)$, and the past output is represented by $y(t-1)$. The symbol $f(x)$ represents the function that models this relation. The nonlinearity introduced by the basic hyperbolic tangent function $f(x)$ in an RNN is due to the fact that it often has unstable gradients. The application of the rectified linear unit function or any other non-saturating activation function leads to arbitrary big gradients. Fig. 6 shows that this model made use of a thick layer in addition to two recurrent layers.

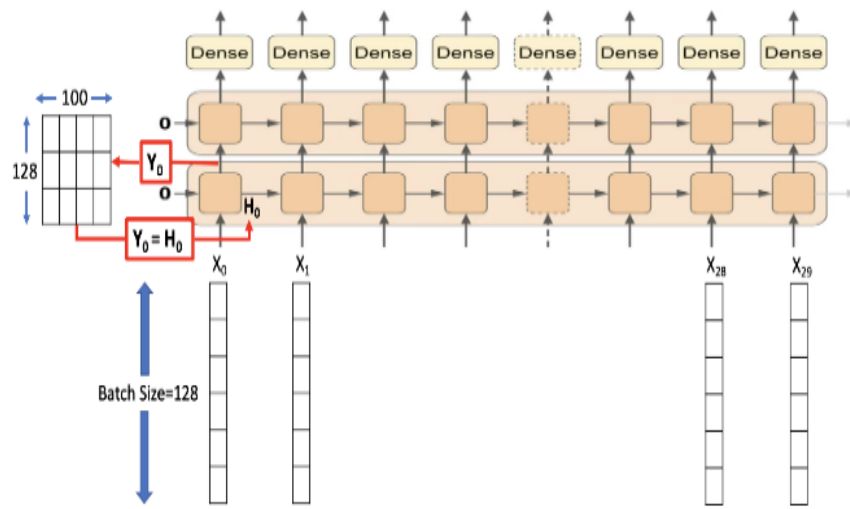


Fig. 6. RNN Model Structure

There are 30 time steps in total, and there are 100 units in each of the recurrent layers. The batch size that may be used is 128. After taking in data from the recurrent layers, the dense layer predicts the total number of questions with a single value. Recurrent neural network (RNN) units in this design store and transmit state information to subsequent iterations. There are a few drawbacks to stateful RNN models: they take more time to train, and they can't always handle longer sequences well due to strong correlations across batches. A 100-epoch, 64-batch RNN was trained using the Adam optimizer with a 0.001 learning rate. Recurrent layers used Tanh activation, and a Dense output layer with 1 neuron was used for insurance premium regression; the network included 100 recurrent units and 30 time steps. When determining model parameters, Mean Squared Error (MSE) was utilized.

Evaluation Metrics

Model evaluation offers a clear assessment of the prediction performance of the suggested models, which makes it a crucial step in machine learning. Evaluating how well the projected values match the actual data is necessary since getting precise predictions in regression tasks is not always easy. Here, we assess the efficacy of the models using four popular regression metrics: R-squared (R^2), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE). These evaluation measures' corresponding equations are shown in Equation (3) to (6) below:

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (3)$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^p)^2} \quad (4)$$

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - y_i^p)^2 \quad (5)$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_i^p| \quad (6)$$

Where:

n is the total number of observations.

y_i is the actual Real Estate Prices at time i .

y_i^p is the predicted Real Estate Prices at time i .

Model performance is shown by a greater value of R^2 , which quantifies the proportion of the dependent variable's variability explained by the model. The root-mean-squared error (RMSE) measures how far predictions deviate from the real values and is shown against the mean-squared error (MSE). The mean squared error (MSE) is a measure of the accuracy of a prediction system; smaller numbers indicate better accuracy. When calculating MAE, the direction of errors is not taken into account; instead, the average absolute difference is measured. More precise forecasts are shown by a lower MAE number. Taken together, these measures offer a thorough assessment of regression efficiency.

Results and Discussion

The recommended Recurrent Neural Network (RNN) for insurance premium prediction is detailed here, along with the

experimental setup and performance evaluation. The Windows 11 Pro operating system, together with a 64 GB DDR5 RAM, an Intel Core i9-13900K CPU (3.0 GHz), and an NVIDIA RTX 4090 GPU (24 GB VRAM), were utilized for the research. Python modules such as Pandas, NumPy, Matplotlib, Seaborn, and Scikit-learn were also utilized. R-Squared, RMSE, MSE, and MAE were used to assess the proposed model's regression performance after training and testing it on the Insurance Premium Prediction Dataset. Table II displays the results of the suggested RNN's predictive performance for insurance premium estimation, which was 92.94% R-Squared, 0.27 RMSE, 0.0727 MSE, and 0.143 MAE.

TABLE II. Classification Results of the Proposed Insurance Premium Prediction Model Using Health Attributes

Matrix	Proposed Recurrent Neural Network (RNN)
R-Squared	92.94
RMSE	0.27
MSE	0.0727
MAE	0.143

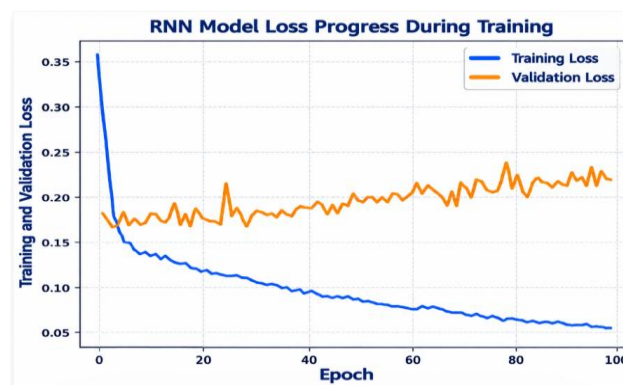


Fig. 7. Rnn Model Loss Progress During Training

The RNN model's training and validation loss across 100 epochs is shown in Fig. 7. The optimization and learning process in training continues as the training loss decreases from 0.36 to 0.05. The validation loss, on the other hand, ranges from about 0.17 to 0.24 and starts to rise at later epochs. The possibility of overfitting happening during training is raised as the difference between the training loss and validation loss grows.

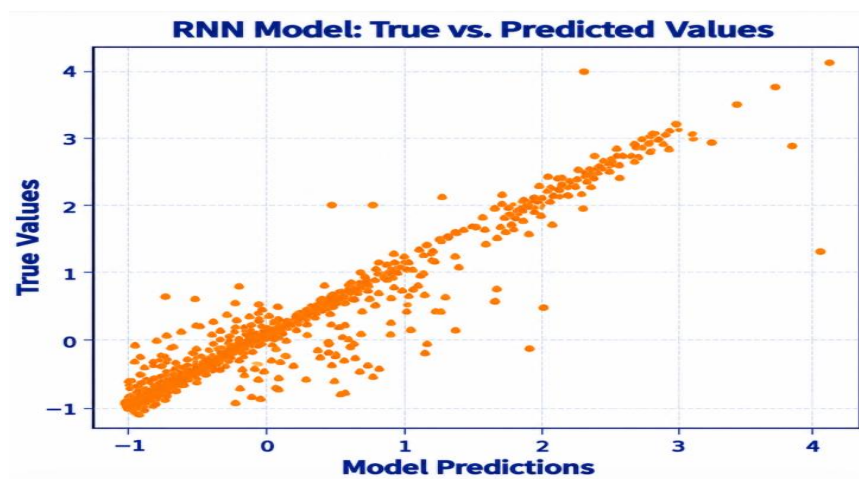


Fig. 8. RNN Model: True Vs. Predicted Values

The correlation between the actual and anticipated target values is shown in Fig 8's scatter plot, which spans from -1 to 4. The majority of the data points are clustered around the diagonal line, suggesting a significant positive correlation and effective prediction. A large number of observations fall within the range of -1 to 1, showing that the model consistently performs well in this lower value range. A small number of dots are off-diagonal, indicating small prediction errors for larger values. As a whole, the model's predicted performance over the observed target range is accurate and dependable.

Comparative Analysis

Table III provides a performance assessment comparison with various current models to help evaluate the proposed model's efficacy. The outcomes analyses the suggested Recurrent Neural Network (RNN) with Linear Regression (LR), Support Vector Regression (SVR), Random Forest (RF), and R-Squared metrics, as well as RMSE, MSE, and MAE. Results for the suggested RNN were as follows: MAE=0.143, R-Squared=92.94%, RMSE=0.27, and MSE= 0.0727. The RNN consistently outperforms the other models when it comes to estimating health insurance premiums. Findings like these show that the suggested model does a good job of capturing connections between health characteristics and insurance premium amounts.

TABLE III. Comparison of Different Machine Learning Models for Health-Based Insurance Premium Prediction

Model	R-Squared	RMSE	MSE	MAE
XGBoost [28]	77.3	5763.70	-	-
LR[29]	78.33	-	210.43	4,186.51
SVR[30]	81	522.6	-	370.9
RF[31]	86.4	0.4592		0.2551
Proposed RNN	92.94	0.27	0.0727	0.143

The suggested RNN has a high R-Squared value of 92.94%, which means it can accurately forecast insurance premium values. It can discover intricate correlations between demographic and health-related characteristics because to its sequential learning capacity. Prediction performance and low error values of the examined metrics are also consistently maintained by the model. When it comes to predicting insurance premiums, the suggested RNN is both effective and dependable.

Conclusion and Future Study

Health insurance coverage helps alleviate the financial strain that comes with illness by covering medical bills. In order to determine the right coverage and enable good insurance planning, it is vital to accurately anticipate premiums, as healthcare and insurance costs are influenced by a variety of demographic and health-related factors. Through the detection of trends in healthcare spending and the prediction of future expenditures, machine learning may enhance the effectiveness of insurance systems. R^2 values of 77.3%, 78.33%, 81.0%, and 86.4% were obtained by XGBoost, Linear Regression (LR), Support Vector Regression (SVR), and Random Forest (RF), respectively, in this investigation. The suggested RNN, however, acquired the greatest R^2 value of 92.94%. These findings show that the suggested RNN is a good method for data-driven insurance cost estimating and that it well captures the correlations between health and demographic characteristics for the purpose of predicting health insurance premiums.

There are just a handful of demographic and health-related variables included in the Insurance Premium Prediction Dataset. To enhance model robustness, generalization, and real-world application, future work can delve into more extensive and varied datasets that incorporate additional medical, lifestyle, and socioeconomic aspects. It can also investigate advanced deep learning and hybrid optimization approaches.

References

1. N. R. Möllmann, M. Mirbabaie, and S. Stieglitz, "Is it alright to use artificial intelligence in digital health? A systematic literature review on ethical considerations," *Health Informatics J.*, vol. 27, no. 4, Oct. 2021, doi: 10.1177/14604582211052391.
2. N. K. S. Patel, "Enterprise AI Operations in Healthcare IT Infrastructure," *Int. J. Intell. Syst. Appl. Eng.*, vol. 14, no. 1s, pp. 2325–2338, Aug, 2026.
3. V. Dogra, A. Singh, S. Verma, Kavita, N. Z. Jhanjhi, and M. N. Talib, "Analyzing DistilBERT for Sentiment Classification of Banking Financial News," in *Lecture Notes in Networks and Systems*, vol. 248, 2021, pp. 501–510. doi: 10.1007/978-981-16-3153-5_53.

4. G. N. Ahmad, H. Fatima, S. Ullah, A. Salah Saidi, and Imdadullah, "Efficient Medical Diagnosis of Human Heart Diseases Using Machine Learning Techniques With and Without GridSearchCV," *IEEE Access*, vol. 10, pp. 80151–80173, 2022, doi: 10.1109/ACCESS.2022.3165792.
5. S. Pujari et al., "Artificial intelligence for global health: cautious optimism with safeguards," *Bull. World Health Organ.*, vol. 101, no. 06, pp. 364-364A, Jun. 2023, doi: 10.2471/BLT.23.290215.
6. K. Hanson et al., "The Lancet Global Health Commission on financing primary health care: putting people at the centre," *Lancet Glob. Heal.*, vol. 10, no. 5, pp. e715–e772, May 2022, doi: 10.1016/S2214-109X(22)00005-5.
7. B. Meessen, "The Role of Digital Strategies in Financing Health Care for Universal Health Coverage in Low- and Middle-Income Countries," *Glob. Heal. Sci. Pract.*, vol. 6, no. Supplement 1, pp. S29–S40, Oct. 2018, doi: 10.9745/GHSP-D-18-00271.
8. H. Kaiser, N. Rotigliano, S. Flessa, B. Ekman, and J. Sundewall, "Extending universal health coverage to informal workers: A systematic review of health financing schemes in low- and middle-income countries in Southeast Asia," *PLoS One*, vol. 18, no. 7, p. e0288269, Jul. 2023, doi: 10.1371/journal.pone.0288269.
9. D. McIntyre, A. G. Obse, E. W. Barasa, and J. E. Ataguba, "Challenges in Financing Universal Health Coverage in Sub-Saharan Africa," in *Oxford Research Encyclopedia of Economics and Finance*, Oxford University Press New York, NY, 2018. doi: 10.1093/acrefore/9780190625979.013.28.
10. Chu, S. Kwon, and P. Cowley, "Health Financing Reforms for Moving towards Universal Health Coverage in the Western Pacific Region," *Heal. Syst. Reform*, vol. 5, no. 1, pp. 32–47, Jan. 2019, doi: 10.1080/23288604.2018.1544029.
11. 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.
12. 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, Research Square. doi: 10.21203/rs.3.rs-10362812/v1.
13. N. Malaviya and S. B. Gogia, "Development, implementation and benefits of a rheumatology-specific electronic medical record application with automated display of outcome measures," *Int. J. Rheum. Dis.*, vol. 13, no. 4, pp. 347–60, Oct. 2010, doi: 10.1111/j.1756-185X.2010.01551.x.
14. W. L. Ho, J. Ali, and K. Caals, "Ensuring trustworthy use of artificial intelligence and big data analytics in health insurance," *Bull. World Health Organ.*, vol. 98, no. 4, pp. 263–269, Apr. 2020, doi: 10.2471/BLT.19.234732.
15. Thesmar, D. Sraer, L. Pinheiro, N. Dadson, R. Veliche, and P. Greenberg, "Combining the Power of Artificial Intelligence with the Richness of Healthcare Claims Data: Opportunities and Challenges," *Pharmacoeconomics*, vol. 37, no. 6, pp. 745–752, Jun. 2019, doi: 10.1007/s40273-019-00777-6.
16. R. A. Sowah et al., "Decision Support System (DSS) for Fraud Detection in Health Insurance Claims Using Genetic Support Vector Machines (GSVMs)," *J. Eng.*, vol. 2019, pp. 1–19, Sep. 2019, doi: 10.1155/2019/1432597.
17. S. Panda, B. Purkayastha, D. Das, M. Chakraborty, and S. K. Biswas, "Health Insurance Cost Prediction Using Regression Models," in *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)*, IEEE, May 2022, pp. 168–173. doi: 10.1109/COM-IT-CON54601.2022.9850653.
18. M. Jödicke et al., "Prediction of health care expenditure increase: how does pharmacotherapy contribute?," *BMC Health Serv. Res.*, vol. 19, no. 1, p. 953, Dec. 2019, doi: 10.1186/s12913-019-4616-x.
19. T. Panch, H. Mattie, and R. Atun, "Artificial intelligence and algorithmic bias: implications for health systems," *J. Glob. Health*, vol. 9, no. 2, Dec. 2019, doi: 10.7189/jogh.09.020318.
20. T. T, S. H. T, V. K. V, and K. R, "Medical Insurance Cost Analysis and Prediction using Machine Learning," in *2023 International Conference on Innovative Data Communication Technologies and Application (ICIDCA)*, IEEE, Mar.

2023, pp. 113–117. doi: 10.1109/ICIDCA56705.2023.10100057.

21. V. Mehta, A. Panchal, K. Shah, and S. Purohit, "Medical Insurance Pricing and Affordability: Patterns, Risks, Cost Drivers," in 2026 2nd International Conference on Big Data & Machine Learning (ICBDML), IEEE, Feb. 2026, pp. 1–6. doi: 10.1109/ICBDML68582.2026.11544600.
22. K. Mishra, K. Mishra, L. Kumar, K. Kumar, and A. Kumar, "Enhancing the Forecasting of Medical Insurance Premiums Through the Use of Multiple Methods," in 2026 International Conference on Intelligent Systems in Engineering, Secured Systems and Cybersecurity (ICISESSC), IEEE, Apr. 2026, pp. 1022–1026. doi: 10.1109/ICISESSC68634.2026.11542626.
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. M. R and M. Arunraj, "Enhancing Healthcare Insurance Cost Prediction with a Metaheuristic Optimization Approach: Comparative Analysis of Machine Learning Models," in 2025 3rd International Conference on Inventive Computing and Informatics (ICICI), IEEE, Jun. 2025, pp. 1225–1232. doi: 10.1109/ICICI65870.2025.11069726.
25. R. Kandula, S. Kalyanapu, S. N. Rayapalli, K. Rao Veerabathina, V. Modugumudi, and S. R. Kanikella, "Medical Insurance Predictive Modelling: An Analysis of Machine Learning Methods," in 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), IEEE, Mar. 2024, pp. 1–5. doi: 10.1109/IATMSI60426.2024.10502643.
26. R. Mary Chittilappilly, S. Suresh, and S. Shanmugam, "A Comparative Analysis of Optimizing Medical Insurance Prediction Using Genetic Algorithm and Other Machine Learning Algorithms," in 2023 International Conference on Advances in Computing, Communication and Applied Informatics (ACCAI), Chennai, India: IEEE, May 2023, pp. 1–6. doi: 10.1109/ACCAI58221.2023.10199979.
27. K. Bhatia, S. S. Gill, N. Kamboj, M. Kumar, and R. K. Bhatia, "Health Insurance Cost Prediction using Machine Learning," in 2022 3rd International Conference for Emerging Technology (INCET), IEEE, May 2022, pp. 1–5. doi: 10.1109/INCET54531.2022.9824201.
28. K. Tanusha, M. Varsha, E. Naveen Kumar, K. Pavan Kalayan, and P.M. Suresh, "Medical Insurance Price Prediction Using ML," *Int. Res. J. Adv. Eng. Manag.*, vol. 3, no. 05, pp. 2111–2116, 2025, doi: 10.47392/irjaem.2025.0333.
29. S. Tejaswi, D. S. Sirisha, D. L. Likhita, S. Ganesh, and C. H. Nirmala, "Prediction of Health Insurance Premiums Using ML," vol. 11, no. 11, pp. 4571–4576, 2025.
30. M. Shanif et al., "Premiumpulse : AI-driven Insurance Cost Modeling," *Int. J. Eng. Res. Technol.*, vol. 15, no. 07, pp. 1–6, July, 2026, doi: 10.5281/zenodo.21588600.
31. S. Gangane, S. Khadse, Y. Wahane, K. Durugkar, S. Sheikh, and P. R. Bambodkar, "Health Insurance Premium Prediction System Using Machine Learning," *Int. J. Sci. Res. Eng. Manag.*, vol. 10, no. 04, pp. 1–9, 2026, doi: 10.55041/ijrsrem61098.