

Deep Learning-Based Medical Insurance Cost Prediction and Analysis

Raman Kumar

Assistant Professor

Department of Computer Science and Applications

Mandsaur University, Mandsaur

raman.kumar@meu.edu.in

Abstract: The rapid advancement of technology is expanding its application possibilities across various human activities. In the healthcare sector, technologies such as the Internet of Things (IoT) support numerous applications, including disease analysis, smart monitoring, survival prediction, and medical cost estimation. Health insurance is now necessary to cover the costs of medical diagnostic and hospital treatment due to the growing cost of healthcare. Consequently, for efficient healthcare planning, it is crucial to precisely forecast individual insurance costs according to demographics, health, lifestyle, and family-related aspects. Existing methods often provide limited prediction accuracy and may not adequately capture complex relationships among these factors. The proposed framework considered a Medical Insurance Cost Dataset containing 1,338 records and 7 attributes and applied data preprocessing, label encoding, Min-Max normalization, SMOTE balancing, and a 70:30 train-test split. Artificial Neural Network (ANN) and Gated Recurrent Unit (GRU) models were developed for medical insurance cost prediction. The proposed GRU achieved an R^2 of 98.91%, MAE of 222.9553, RMSE of 865.2828, and MAPE of 1.0952, while the ANN achieved an R^2 of 97.94%, MAE of 508.2767, RMSE of 1156.8298, and MAPE of 2.1902. The results show that the suggested DL framework is capable of accurate and dependable healthcare insurance cost prediction.

Keywords: Medical Insurance, Cost Prediction, Deep Learning, Healthcare Analytics, Healthcare, Medical Cost Estimation, Predictive Modeling.

I. INTRODUCTION

Healthcare is a fundamental component of human well-being and plays an essential role in maintaining a healthy and productive life. Unexpected diseases, chronic diseases, accidents, and medical emergencies can create substantial financial burdens for individuals and families[1][2]. The increasing cost of diagnosis, treatment, hospitalization, medication, disease management, and preventive care has made access to affordable healthcare a major concern. Medical insurance provides financial protection by covering or reducing healthcare-related expenses through reimbursement or cashless treatment. Therefore, an effective medical insurance system is important for protecting individuals from unexpected medical expenditures and maintaining long-term financial stability[3][4][5].

Medical insurance costs vary among individuals because they are influenced by demographic, health, lifestyle, and socioeconomic factors[6][7]. Attributes such as age, gender, BMI, smoking habits, number of children, geographic region, and health conditions can affect medical expenditure and insurance premiums[8]. Accurate cost estimation helps insurance providers develop appropriate pricing strategies, manage financial risks, and allocate healthcare resources efficiently[9][10][11][12][13].

The increasing complexity and volume of healthcare data have created a need for effective methods for predicting and analyzing medical insurance costs[14]. Traditional statistical and ML approaches may face difficulties in identifying complex relationships among demographic, lifestyle, and health-related factors[15][16][17]. Deep Learning (DL)



methods can learn nonlinear patterns and interactions from healthcare data, providing an effective approach for medical insurance cost prediction and analysis[18][19]. In this context, ANN and GRU can be applied to model insurance expenditure and evaluate prediction performance. Accurate cost prediction and analysis can support insurance planning, premium assessment, financial risk management, and healthcare resource allocation[20][21]. DL algorithms may extract complicated patterns from huge, multidimensional healthcare datasets. The integration of ML and DL can support accurate, efficient, and data-driven medical insurance cost prediction and analysis[22][23].

A. Motivation and Contribution

The motivation behind this study lies in the rising cost of healthcare places a significant financial burden on individuals and families, creating a need for accurate medical insurance cost prediction. Insurance expenses depend on multiple demographic, health, lifestyle, and regional factors, making cost estimation challenging. Traditional approaches may have limitations in capturing complex and nonlinear relationships among these factors. Therefore, this study is motivated to develop ANN and GRU-based models for efficient, reliable, and data-driven medical insurance cost prediction and analysis. This research offers several key contributions as listed below:

- Developed a machine learning-based framework for medical insurance cost prediction using demographic, health, and lifestyle attributes.
- Utilized a Medical Insurance Cost Dataset containing 1,338 records and 7 attributes for systematic analysis and prediction.
- Applied data preprocessing techniques, including missing-value handling, label encoding, Min-Max normalization, and consistency checking.
- Used SMOTE (Synthetic Minority Over-sampling Technique) to address class imbalance and improve data representation.
- Implemented two DL models, ANN and Gated Recurrent Unit (GRU), for insurance cost prediction.
- Evaluated prediction performance using R^2 , MAE, RMSE, and MAPE metrics.
- Provided a data-driven approach that can support insurance cost estimation, healthcare planning, and expense analysis.

B. Justification and Novelty

The requirement for precise medical insurance cost projection based on lifestyle, health, and demographic characteristics is addressed by the suggested approach. It is a combination of systematic preprocessing, normalization and data balancing that enhances data quality and model learning. The novelty lies in employing ANN and GRU models to capture complex nonlinear relationships among insurance-related attributes. The framework offers a confidence-inspiring data-driven resource for estimating healthcare expenses and organizing insurance.

C. Structure of the Proposed Study

The organization of the paper is as follows: Section II discusses existing research on medical insurance cost prediction and analysis. Section III describes the dataset, preprocessing techniques, and model development. Section IV presents the experimental results and comparative performance analysis, while Section V summarizes the major findings, limitations, and potential directions for future research.

II. LITERATURE REVIEW

The literature review highlights datasets, predictive models, assessment metrics, conclusions, constraints, and new research prospects as it looks at modern ML and DL techniques for medical insurance cost prediction.

Udayagiri and V, (2026) evaluate the efficacy of attention-based models in comparison to conventional baseline models for forecasting medical insurance costs. They used 1338 de-identified health insurance information to test five models: LinearRegression, RF, XGBoost, LightGBM, and CatBoost, as well as TabNet and a Tabular transformer. They used



the root-mean-squared error and the R² to assess how well it worked. A notable finding from the analysis of the results is that about 60.69% of the total significance is due to smoking-related factors[24].

Jaishree, Kokila and Hegde, (2025) provides an ML-based system that uses lifestyle and demographic characteristics, including age, gender, BMI, number of children, SmokingHabits, and residence, to forecast medical insurance costs. The Kaggle dataset was preprocessed and used to train LinearRegression, DT, RF, GB, and XGBoost models. Performance was evaluated using R², MSE, and RMSE, with XGBoost achieving the highest R² of 0.9548. The best-performing model was deployed through a Flask web application for real-time cost estimation. The work further incorporated SHAP-based interpretability, providing practical usability and transparent predictions beyond accuracy-focused approaches[25].

R and Arunraj, (2025) propose a metaheuristic optimization algorithm for hyperparameter tuning and feature selection. Benchmark methods including Genetic Algorithm (GA), PSO, and Bayesian Optimization were compared using MAE, MSE, RMSE, and R². The proposed algorithm applied to RF and GB achieved the best performance, reaching an R² score of 92.5% and an RMSE of 2510.23, improving accuracy by 4.3%. Ensemble-based models outperformed regression-based models in predicting insurance costs[26].

Mohan et al., (2025) uses a ML algorithm with prospective health risk variables such as age, sex, BMI, SmokingStatus, and city to predict the cost of health insurance claims. A health insurance dataset obtained from Kaggle was used, involving DataPreprocessing, FeatureEngineering, and model selection with R² and MAE as evaluation measures. The outcomes indicated that RF achieved the best performance among the tested models, with an accuracy of 96.7% in modeling the health data[27].

Vuddanti et al., (2024) experienced numerous health problems during the COVID-19 pandemic. The creation and assessment of predictive models for the purpose of calculating medical costs are the subjects of investigation. This study has utilized the 4968-row USA medical insurance dataset from Kaggle. The dataset's features allow a summary of an individual's personal and medical circumstances. Regression analysis and gradient boosting are the two machine learning methods utilized to build predictive models. The analysis reflected the link between features and costs. After training the model, a 94% accuracy rate has been obtained[28].

R. Mary Chittilappilly, S. Suresh, and S. Shanmugam, (2023) examined several ML techniques for predicting health insurance premiums based on individuals' demographic and health data. Each algorithm was trained and tested on a preprocessed dataset, and the performance of various models is compared, including LinearRegression, RF, GB, LassoLarsCV, and a model created with automated machine learning using TPOT, which makes use of genetic programming for optimization. Conclusively, the results bring to light that the TPOT-generated model, which is a combination of LassoLarsCV and GradientBoostingRegressor, performed better than the other models, attaining a root mean square error of 0.0686 and an accuracy of 87.45% on the test set[29].

The evaluated research mostly concentrates on explainability, optimization, ensemble approaches, and conventional ML for medical insurance cost prediction. While transformer and attention-based models have been investigated, limited research has explored DL architectures for modelling interactions among lifestyle, health, and demographic factors. Additionally, comparative analysis of DL models using consistent preprocessing and multiple regression metrics remains limited. Recent research is summarized in Table I, highlighting methodologies, datasets, key findings, limitations, and opportunities for developing advanced DL-based approaches.

Table I Recent Research on Medical Insurance Cost Prediction and Analytical Methods

Author	Proposed Work	Dataset	Key Findings	Limitations & Future Work
Udayagiri and V. (2026)	Compared conventional, ensemble, attention-based, and transformer models for insurance cost forecasting.	1,338 insurance records	Smoking-related factors contributed approximately 60.69% of feature significance.	Deep learning-based cost prediction can be further investigated with broader feature analysis.



Jaishree, Kokila and Hegde (2025)	Developed ML models with SHAP-based interpretability for insurance cost estimation.	Kaggle medical insurance dataset	XGBoost achieved the highest R^2 of 0.9548 with interpretable predictions.	Advanced DL models and larger datasets can be explored to improve prediction.
R and Arunraj (2025)	Applied metaheuristic optimization for feature selection and hyperparameter tuning.	Medical insurance dataset	Optimized ensemble models achieved R^2 of 92.5% and RMSE of 2510.23.	More robust DL-based optimization and generalization studies are needed.
Mohan et al. (2025)	Used ML algorithms to estimate insurance claim costs from health-risk factors.	Kaggle health insurance dataset	Random Forest achieved the best reported performance with 96.7% accuracy.	DL approaches and additional healthcare attributes can be evaluated.
Vuddanti et al. (2024)	Investigated regression and gradient boosting for medical cost prediction.	4,968-row Kaggle dataset	The models examined relationships between individual characteristics and medical costs.	More advanced predictive models and comprehensive validation are required.
R. Mary Chittilappilly, S. Suresh, and S. Shanmugam (2023)	Compared ML and automated ML approaches for medical insurance charge prediction.	Preprocessed medical insurance dataset	TPOT-based modeling achieved RMSE of 0.0686 and 87.45% reported accuracy.	DL architectures and larger, diverse datasets remain to be explored.

Research gaps: Existing studies primarily focus on conventional regression and ensemble machine learning models, with limited exploration of deep learning architectures for predicting medical insurance costs. The majority of methods focus on achieving high prediction accuracy, and do not investigate complex non-linear relationships between the demographic, health, and lifestyle variables. Additionally, the limited comparative analysis between ANN and sequential learning architectures like GRU was carried out based on multiple regression evaluation metrics. Therefore, an integrated deep learning framework is needed to improve prediction reliability and provide comprehensive medical insurance cost analysis.



III. RESEARCH METHODOLOGY

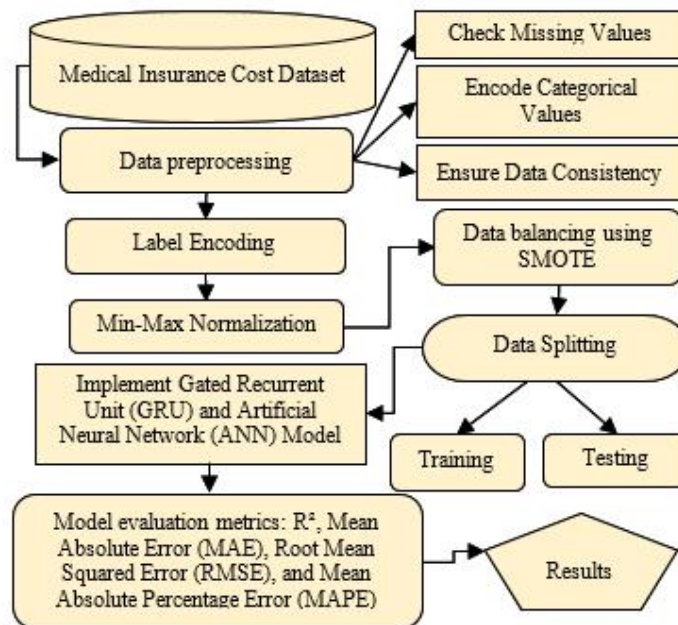


Fig. 1 Proposed Framework for Medical Insurance Cost Prediction and Analysis

The proposed methodology includes the Medical Insurance Cost Dataset (1338 records and 7 attributes), data cleaning, label encoding, Min-Max normalization, SMOTE balancing, and train-test split (70:30). ANN and GRU models are trained on the preprocessed data to learn complex relationships among demographic, health, and insurance-cost features. Methods for assessing the model's efficacy in making accurate predictions include R^2 , MAE, RMSE, and MAPE. The flowchart in Fig. 1 shows the advanced medical insurance cost prediction and analysis process.

The following section provides a detailed description of each step involved in the proposed methodology:

A. Data Collection and Analytical Assessment

The Medical Insurance Cost Dataset includes 1,338 records and 7 characteristics for cost prediction and analysis, such as age, gender, BMI, children, smoking status, area, and medical insurance costs. The collected attributes were analyzed to identify patterns and relationships between demographic, health, and insurance-cost factors. Data visualizations such as bar plots and heatmaps were used to examine cost distributions and feature correlations, as shown below:

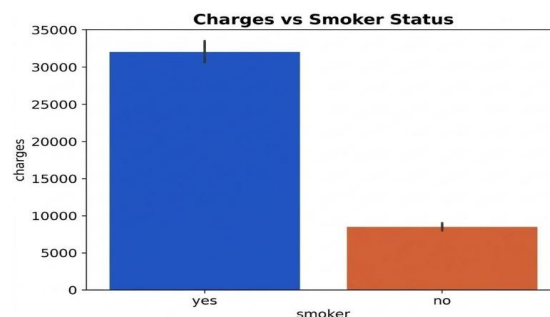


Fig 2 Bar Graph of Class Distribution



Fig 2 illustrates a bar graph that compares the average insurance charges by smoking behaviours. The graph indicates that the costs of smokers are substantially higher than those of non-smokers, with a difference of approximately 32,000 and 8,500. The dark vertical error bars atop each bar indicate the confidence intervals and variability around these calculated mean charges. Overall, the chart clearly highlights smoking status as a major driving factor in determining individual healthcare expenses.

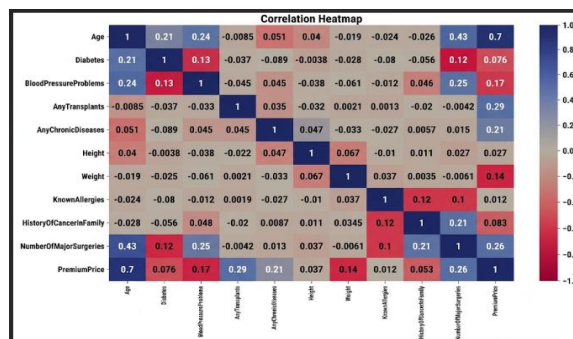


Fig 3 Correlation Matrix Heatmap

Fig 3 illustrates the pairwise linear correlation coefficients among key health metrics and demographic features, including age, medical conditions, and insurance pricing. The intensity of the colour in the matrix grid distinguishes between stronger positive relationships, for example between Age and PremiumPrice (0.7). If the numbers are close to zero, the features are approximately linearly independent, such as Height vs Weight, or medical history vs. other medical history variables. It is a crucial matrix to determine feature redundancy and the most important predictors before training the model.

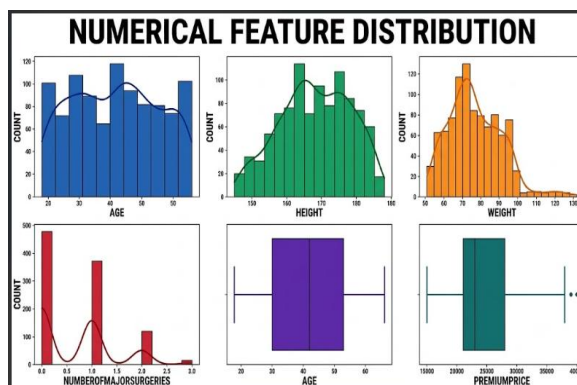


Fig 4 Numerical Feature Distribution

Fig 4 visualization shows 6 subplots showing the spread, the center and the skewness of the data for each of the important metrics. The top row shows histograms with kernel density estimation (KDE) curves highlighting a fairly uniform distribution for Age, a near-normal distribution for Height, and a right-skewed pattern for Weight. Finally, horizontal box plots are provided for the variables Age and PremiumPrice in the bottom row, with the variable NumberOfMajorSurgeries showing a heavy right skew and corresponding horizontal box plot.

B. Data Pre-processing

The Medical Insurance Cost Dataset was preprocessed by handling missing values, encoding categorical features, and applying Min-Max normalization to scale numerical values between 0 and 1. SMOTE was then utilized to balance the dataset, preparing the data for effective ML model training and evaluation. The main preprocessing procedures are summarized as follows:



- **Check Missing Values:** Missing or incomplete records were identified and appropriately handled to maintain dataset quality.
- **Encode Categorical Values:** Categorical characteristics were changed to numerical values in order to be compatible with ML methods.
- **Ensure Data Consistency:** The cleaned dataset was checked for consistency and completeness before further preprocessing and modelling.

C. Label Encoding

The dataset contains categorical attributes that cannot be directly interpreted by most machine learning algorithms. Therefore, label encoding was used to transform category data into numerical representations. For example, binary categorical attributes were assigned the numerical values 0 and 1, while the remaining categorical features were similarly converted to numerical values.

D. Min-Max Normalization

Min-Max scaling was applied to transform NumericalFeatures into a common range between 0 and 1. This prevents features with larger numerical values from dominating the learning process and improves consistency among input variables. The scaling was performed using 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

E. Data balancing using SMOTE

Data balancing is the process of reducing class imbalance by increasing the representation of underrepresented classes in the dataset. In this study, Synthetic Minority Over-sampling Technique (SMOTE) was applied to generate synthetic samples for the minority class. SMOTE identifies the nearest neighbors of each minority-class sample and creates new instances through interpolation between the sample and its neighbors. The generated samples are added to the training data, resulting in a more balanced ClassDistribution and helping the ML model learn minority-class patterns more effectively.

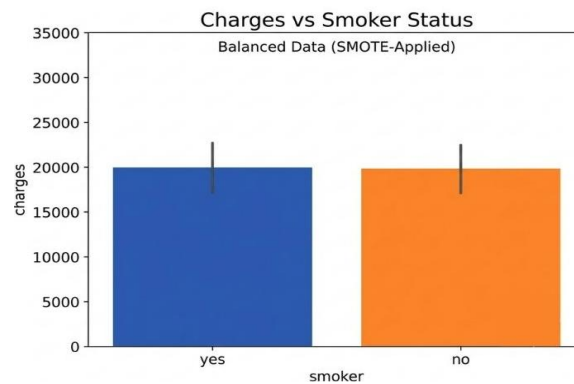


Fig. 5 Bar Graph of Class Distribution After SMOTE

Fig. 5 illustrates insurance charges across smoking categories following the application of SMOTE oversampling. The rebalanced distributions show stabilized average costs for both "yes" and "no" smoker groups centered around 20,000. Dark error bars on top of each bar show confidence bounds assuming equal number of samples per bar, equal class representation.

F. Data Splitting

In machine learning modeling, dataset splitting is an essential step that aids in various stages, from training to evaluating the model. There are two separate subsets of the dataset: a testing set and a training set. 30% of the data is utilized for testing, and 70% for training throughout the experimental phase.



G. Models Implementation

The proposed ANN and GRU models were implemented to predict and analyse medical insurance costs using demographic and health-related attributes. Both models were trained on the preprocessed dataset to learn complex relationships among features and improve predictive performance.

H. Proposed Artificial Neural Network (ANN) Model

ANN is a non-linear computational model that has played a vital role in a field of ML. It is based on the working mechanism of a biological neural network. Similar to the linked neurons of the human brain, an ANN is made up of a large number of linked neurons. These neurons are capable of learning, generalizing training data, and deriving conclusions from complex data. The use of ANNs in the construction industry for risk management has gained attention. ANNs, inspired by the structure and cognitive processes of the human brain, provide a powerful computational model for pattern recognition and complex data processing. These networks mimic neuronal interactions in the brain, a concept that has been extensively explored in the literature, including seminal works by scholars like Ian Goodfellow. ANNs' capacity to learn intricate, non-linear correlations between features is one of its main advantages. This contrasts with linear regression models, as ANNs can autonomously learn feature transformations, leading to more precise predictions. This is especially valuable in the construction industry, where data is often multifaceted and unpredictable. Additionally, ANNs are known for their flexibility and ease of use, making them suitable for uniform application across diverse tasks.

ANN is selected as the uniform model used for cyber risk assessment in construction. A feedforward ANN, composed of neurons in multiple layers, utilizes hidden layers to perform this automatic transformation. The operation of a hidden layer with m neurons, following a layer with n neurons, is described by Equation (2).

$$h_j = \sigma(\sum_{i=1}^n w_{ij}x_i + b_j) \quad (2)$$

Where,

h_j —the value of the j th neuron in the hidden layer,

x_i —the value of the i th neuron in the previous layer,

w_{ij} —the weight of the j th neuron in the previous layer that connects to the j -th neuron in the hidden layer,

b_j —the bias term for the j th neuron in the hidden layer,

σ —the activation function.

The proposed ANN uses 2 hidden layers with 64 and 32 neurons, ReLU activation, Adam optimizer, learning rate 0.001, batch size 32, and 100 epochs.

I. Proposed Gated Recurrent Unit (GRU) Model

The internal structure of an LSTM neural network is more complex, and it is difficult to adjust its parameters, which leads to longer time spent on model training. A simplified version of an LSTM called GRU. The GRU model requires less training time than the LSTM model, offering similar predictive accuracy. GRU integrates the input gate and forget gate of LSTM into a single update gate, reducing the memory module to only two gating components: the update gate and the reset gate. The update gate, represented as Z_t , manages the degree to which previous data state information is incorporated into the current state. A higher value of the update gate indicates that the present neuron retains more historical information, while the previous neuron retains comparatively less. The primary function of the update gate is to refresh the memory and discern long-term patterns in the water quality data sequence. The formula for the update gate to capture information can be expressed as the following in Equation (3)

$$Z_t = \sigma(W_z * [h_{t-1}, X_t]) \quad (3)$$

The reset gate, denoted as R_t , plays a crucial role in assessing the extent of historical information retention. A lower reset gate value signifies a higher retention of past historical data, which is conducive to capturing the short-term patterns in the water quality parametric data. The formula for the reset gate to obtain information is the following in Equation (4)

$$R_t = \sigma(W_r * [h_{t-1}, X_t]) \quad (4)$$



where h_t denotes the output state of the unit at moment t and $\bar{h}_t$ denotes the implied state value of h_t . This value is utilized for storing the data of the present unit and transmitting it to the subsequent unit, while the estimated value for the previous moment's output is computed as Equation (5)

$$\bar{h}_t = \tanh(W_{\bar{h}} * [r_t * h_{t-1}, X_t]) \quad (5)$$

The forecasted outcomes of the water quality parameter data can be expressed as Equation (6)

$$h_t = (1 - Z_t) * h_{t-1} + Z_t * \bar{h}_t \quad (6)$$

where X_t denotes the input data value at the current moment t , $h_{(t-1)}$ is the output value of the water quality parameter data in the memory cell at moment $t-1$, W_Z, W_r , and $W_{\bar{h}}$ denote the weight matrix in the cell, “[]” denotes the connection of two matrices, “*” denotes the matrix product, σ is the activation function, and $\tanh$ is the bisecting curve of the activation function. The proposed GRU uses 2 GRU layers with 64 and 32 units, $\tanh$ activation, Adam optimizer, learning rate 0.001, batch size 32, dropout 0.2, and 100 epochs.

J. Evaluation Metrics

Model assessment is a critical stage in evaluating the model's performance and determining the similarity between anticipated and real insurance prices. In this study, the proposed models were evaluated using four metrics: R^2 , MeanAbsoluteError (MAE), RootMeanSquaredError (RMSE), and MeanAbsolutePercentageError (MAPE). All these metrics together offer a reliable measure of model accuracy, model prediction error and overall model reliability.

1) R-Squared (R^2)

R^2 is a statistical measure used to determine the proportion of variability in the response variable explained by the regression model. Its value may be anywhere from 0 to 1, with 1 being the best possible prediction of the response data and 0 the worst possible explanation of variability. The formula for calculating R^2 is given in Equation (7):

$$R^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y}_i)^2} \quad (7)$$

2) Mean Absolute Error (MAE)

MAE is a widely used statistic for assessing the accuracy of prediction models. It calculates the average size of errors in a group of predictions without considering their direction. A lower MAE value suggests better performance. The formula to calculate MAE is Equation (8):

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - y_i^p| \quad (8)$$

3) Root Mean Squared Error (RMSE)

This metric represents the square root of MSE. RMSE measures the degree to which the model's predictions differ from the actual values. Lower RMSE values indicate better model performance. The formula to calculate RMSE is Equation (9):

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - y_i^p)^2} \quad (9)$$

Mean Absolute Percentage Error (MAPE)

MAPE computes errors in terms of percentage; more specifically, it is the average percentage difference between predictions and their intended targets in the dataset. MAPE can also be seen as the MAE returned as a percentage Equation (10).

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left(\frac{(y_i - y_i^p)^2}{y_i} \right) / * 100 \quad (10)$$

Where,

- y is the actual value,
- y is the predicted value, and n is the number of observations.
- These metrics together give information about the model's accuracy and efficacy in predicting the TargetVariable.



IV. RESULTS AND DISCUSSION

This section describes the experimental setup and presents the performance of the proposed models across both phases, demonstrating their evaluation and efficient processing. The suggested models' experimental findings are summarised in Table II. The experiments were conducted on a Windows 11 system equipped with an Intel Core i7 11th-generation processor, 16 GB RAM, and NVIDIA GTX 1650 GPU. Python libraries such as Pandas, NumPy, and Scikit-learn were used for model development, while R^2 , MAE, RMSE, and MAPE were used for performance evaluation. The proposed Gated Recurrent Unit (GRU) achieved an R^2 of 98.91%, whereas the ANN obtained 97.94%, with corresponding MAE, RMSE, and MAPE values of 222.9553, 865.2828, 1.0952 and 508.2767, 1156.8298, 2.1902, respectively. These results prove that the suggested models are useful for analysing and predicting health insurance premiums.

Table II. Performance Results of the Proposed Models for Medical Insurance Cost Prediction and Analysis

Matrix	Proposed Gated Recurrent Unit (GRU)	Proposed Artificial Neural Network (ANN)
R-Square (R^2)	98.91	97.94
MAE	222.9553	508.2767
RMSE	865.2828	1156.8298
MAPE	1.0952	2.1902

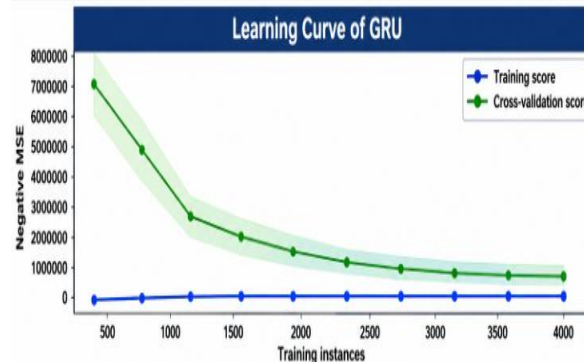


Fig. 6 Learning Curve of GRU

Fig. 6 GRU learning curve shows a sharp reduction in negative MSE as training instances increase. The training and cross-validation scores gradually converge, indicating improved model learning and generalization. The decreasing validation error demonstrates stable performance with larger training samples.

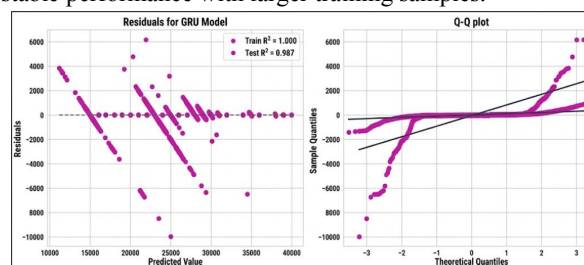


Fig. 7 Residuals and Q-Q Plot for the GRU Model

Fig. 7 depicts the model performance, with the left plot depicting the distribution of residuals over the training and test R^2 with a value of 1.000 and 0.987 respectively, and the right plot a QQ plot comparing the sample quantiles with the theoretical quantiles to assess residual normality.



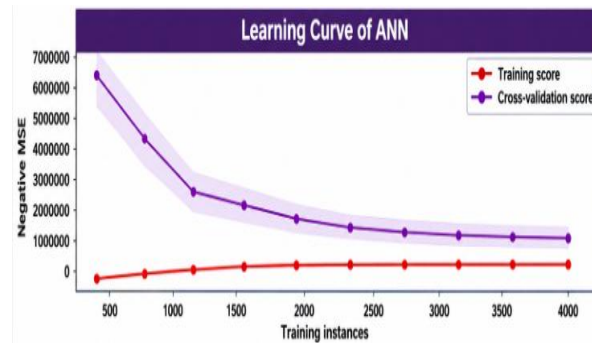


Fig 8 Learning Curve of ANN

In Fig 8 ANN learning curve shows a steady decline in negative MSE with increasing training instances. The training and cross-validation curves become more stable as the sample size increases. This trend indicates improved prediction performance and better generalization of the ANN model.

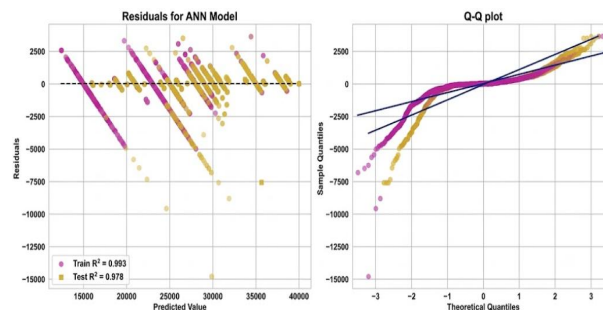


Fig. 9 Residuals and Q-Q Plot for the Artificial Neural Network (ANN) Model

Fig. 9 illustrate performance diagnostics through two side-by-side plots. The left subplot tracks residual scatter across predicted values against the zero-error reference line, reporting a high goodness-of-fit with a Train $R^2 = 0.993$ and Test $R^2 = 0.978$. The right Q-Q plot compares sample quantiles to theoretical quantiles along fitted linear trendlines to assess the normality of the error distribution.

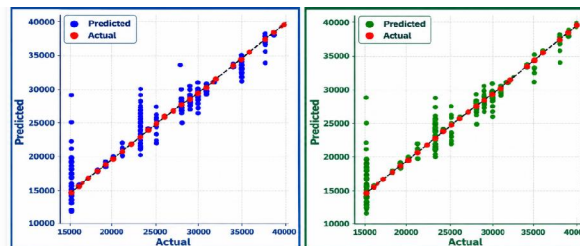


Fig 10 Testing Set: Actual vs. Predicted (GRU and ANN)

Fig 10 compares the actual and predicted test values for the GRU and ANN models, where most predicted points are closely aligned with the diagonal reference line, indicating strong agreement with the actual values. The GRU model has an R^2 value of 98.91% and the ANN model has a value of 97.94%, showing that both models have high prediction accuracy. The proximity of the points suggests that there aren't many discrepancies between the predicted and actual values. Overall, both models demonstrate strong predictive performance and effective generalization on the testing set.

A. Comparative analysis

To assess the effectiveness of the proposed models, a comparative performance evaluation with existing machine learning models is presented in Table III. The results show that KNN, Linear Regression, SVM, and XGBoost achieved R^2 values of 55.21%, 73.57%, 84.72%, and 86.47%, respectively. The proposed ANN achieved an R^2 of 97.94%, while



the proposed GRU obtained the highest R^2 of 98.91%, along with the lowest MAE (222.9553), RMSE (865.2828), and MAPE (1.0952). These findings demonstrate the strong predictive performance of the proposed deep learning models for predicting and analysing medical insurance costs.

Table III. Comparison of Machine Learning and Deep Learning Models for Medical Insurance Cost Prediction and Analysis

Model	R^2	MAE	RMSE	MAPE
KNN[30]	55.21	-	4431.1	-
LR[31]	73.57	4181.7399	6050.4615	-
SVM[32]	84.72	-	0.358771	-
XGBoost[33]	86.470	1442.904	2231.524	5.906
Proposed ANN	97.94	508.2767	1156.8298	2.1902
Proposed GRU	98.91	222.9553	865.2828	1.0952

The proposed ANN and GRU models effectively capture complex relationships among medical and demographic factors. The GRU achieved an R^2 of 98.91%, while ANN achieved 97.94%, demonstrating strong prediction performance. Both models help reduce prediction errors and provide reliable medical insurance cost estimates. Their predictive capabilities can support effective insurance planning, premium assessment, and cost analysis.

V. CONCLUSION AND FUTURE STUDY

Predicting healthcare expenses accurately is crucial for a wide range of stakeholders and healthcare institutions. The healthcare industry generates massive volumes of data related to patient demographics, health conditions, and medical expenses, which can be effectively analyzed to provide valuable insights into insurance costs. Machine learning (ML) techniques can improve healthcare cost prediction by identifying complex relationships among demographic and health-related factors. In this study, the proposed ANN and GRU models achieved R^2 values of 97.94% and 98.91%, respectively, outperforming K-Nearest Neighbors (KNN) (55.21%), Linear Regression (LR) (73.57%), SVM (84.72%), and XGBoost (86.47%). These findings show how well the suggested deep learning models work for data-driven healthcare expenditure analysis and precise medical insurance cost prediction.

The sample size and the omission of some demographic and health-related characteristics may limit the generalizability of this study. Future work can incorporate larger, more diverse, and real-world datasets with additional healthcare and socioeconomic factors to further improve prediction robustness.

REFERENCES

1. C. Yang, C. Delcher, E. Shenkman, and S. Ranka, "Machine learning approaches for predicting high cost high need patient expenditures in health care," *Biomed. Eng. Online*, vol. 17, no. S1, pp. 81–118, Nov. 2018, doi: 10.1186/s12938-018-0568-3.
2. K. Dutta, S. Chandra, M. K. Gourisaria, and H. GM, "A Data Mining based Target Regression-Oriented Approach to Modelling 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.
3. B. D. Sommers, "Health Insurance Coverage: What Comes After The ACA?," *Health Aff.*, vol. 39, no. 3, pp. 502–508, Mar. 2020, doi: 10.1377/hlthaff.2019.01416.
4. B. Milovic, "Prediction and decision making in Health Care using Data Mining," *Int. J. Public Heal. Sci.*, vol. 1, no. 2, Dec. 2012, doi: 10.11591/ijphs.v1i2.1380.
5. P. Chandra Shekhar, "Next-Gen Test Automation in Life Insurance: Self-Healing Frameworks," *Int. J. Sci. Res. Eng. Manag.*, vol. 5, no. 7, pp. 1–13, 2021, doi: 10.55041/IJSREM9554.



6. L. Luo, X. Yu, Z. Yong, C. Li, and Y. Gu, "Design Comorbidity Portfolios to Improve Treatment Cost Prediction of Asthma Using Machine Learning," IEEE J. Biomed. Heal. Informatics, vol. 25, no. 6, pp. 2237–2247, Jun. 2021, doi: 10.1109/JBHI.2020.3034092.
7. Okfalisa, I. Gazalba, Mustakim, and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, Nov. 2017, pp. 294–298. doi: 10.1109/ICITISEE.2017.8285514.
8. G. Marinova and M. Todorova, "Regression Analysis for Predicting Health Insurance," in 2023 4th International Conference on Communications, Information, Electronic and Energy Systems (CIEES), IEEE, Nov. 2023, pp. 1–4. doi: 10.1109/CIEES58940.2023.10378755.
9. T. Chen and C. Guestrin, "XGBoost," in Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, NY, USA: ACM, Aug. 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
10. C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," Math. Probl. Eng., vol. 2021, pp. 1–13, Dec. 2021, doi: 10.1155/2021/1162553.
11. R. D, M. S. K, and D. J, "Health Insurance Cost Prediction using Machine Learning Algorithms," in 2022 International Conference on Edge Computing and Applications (ICECAA), IEEE, Oct. 2022, pp. 1381–1384. doi: 10.1109/ICECAA55415.2022.9936153.
12. D. P. Dāsepalli, "Artificial Intelligence in Digital Banking: A Systematic Literature Review and a Novel CrossDomain Resilience Framework for Customer Experience, Fraud Detection, and Risk Management," Res. Sq., pp. 1–26, July, 2026, doi: <https://doi.org/10.21203/rs.3.rs-10362812/v1>.
13. S. Gupta and D. K. Jayswal, "Enhanced Artificial Intelligence (AI)-Based Risk Prediction Framework for Cardiovascular Health Monitoring," in 2026 2nd International Conference on Artificial Intelligence Systems (AIS), San Antonio, TX, USA: IEEE, May 2026, pp. 282–288. doi: 10.1109/AIS69919.2026.11621536.
14. A. 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.
15. 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.
16. 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.
17. 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, 2022.
18. 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), IEEE, Feb. 2026, pp. 1–6. doi: 10.1109/ICAIC67076.2026.11395709.
19. S. Murumkar and S. Sen, "A Critical Review on Ethical and Privacy Challenges of AI-Powered Healthcare Systems," Int. J. Artif. Intell. Data Sci. Mach. Learn., vol. 6, no. 1, pp. 254–257, March, 2025, doi: 10.63282/3050-9262.IJAIDSML-V6I1P129.
20. A. Cevolini and E. Esposito, "From pool to profile: Social consequences of algorithmic prediction in insurance," Big Data Soc., vol. 7, no. 2, Jul. 2020, doi: 10.1177/2053951720939228.



21. 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.
22. H. Zhang, W. Zhou, and D. Zhang, "Direct Medical Costs of Parkinson's Disease in Southern China: A Cross-Sectional Study Based on Health Insurance Claims Data in Guangzhou City," Int. J. Environ. Res. Public Health, vol. 19, no. 6, p. 3238, Mar. 2022, doi: 10.3390/ijerph19063238.
23. 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, pp. 1–15, Jun. 2022, doi: 10.3390/ijerph19137898.
24. J. Udayagiri and S. R. V, "Attention-Based Deep Learning for Medical Insurance Cost Prediction: A Comparative Analysis of TabNet and Tabular Transformers," in 2026 4th International Conference on Knowledge Engineering and Communication Systems (ICKECS), IEEE, Apr. 2026, pp. 1–11. doi: 10.1109/ICKECS70176.2026.11528016.
25. D. G. Jaishree, R. Kokila, and D. Hegde, "Medical Insurance Cost Forecasting Using Machine Learning Techniques," in 2025 International Conference on Intelligent Communication Networks and Computational Techniques (ICICNCT), 2025, pp. 1–5. doi: 10.1109/ICICNCT66124.2025.11232894.
26. R. Mahaveerakannan and M. Arunraj, "Enhancing Healthcare Insurance Cost Prediction with a Metaheuristic Optimization Approach: Comparative Analysis of Machine Learning Models," in Proceedings of the 2025 3rd International Conference on Inventive Computing and Informatics, ICICI 2025, 2025, pp. 1225–1232. doi: 10.1109/ICICI65870.2025.11069726.
27. 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.
28. 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.
29. 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), IEEE, May 2023, pp. 1–6. doi: 10.1109/ACCAI58221.2023.10199979.
30. G. K. Patra, C. Kuraku, S. Konkimalla, V. N. Boddapati, M. Sarisa, and M. S. Reddy, "An Analysis and Prediction of Health Insurance Costs Using Machine Learning-Based Regressor Techniques," J. Data Anal. Inf. Process., vol. 12, no. 04, pp. 581–596, 2024, doi: 10.4236/jdaip.2024.124031.
31. Anmol, S. Aggarwal, and A. Jahan Badhon, "Medical Insurance Cost Prediction Using Machine Learning Algorithms," in Lecture Notes in Networks and Systems, vol. 467, 2023, pp. 271–281. doi: 10.1007/978-981-19-2538-2_27.
32. C. A. ul Hassan, J. Iqbal, S. Hussain, H. AlSalman, M. A. A. Mosleh, and S. Sajid Ullah, "A Computational Intelligence Approach for Predicting Medical Insurance Cost," Math. Probl. Eng., vol. 2021, 2021, doi: 10.1155/2021/1162553.
33. U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," Mach. Learn. with Appl., vol. 15, no. 21, pp. 1–20, Mar. 2023, doi: 10.1016/j.mlwa.2023.100516.

