iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** VII **Month of publication:** July 2026

DOI: https://doi.org/10.22214/ijraset.2026.84069

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue VII July 2026- Available at www.ijraset.com

Predictive Modeling of Medical Insurance Expenditure Using Machine Learning on Multi-Context Healthcare Data

Amol Chaudhary[1], Mahika Garg[2], Roopali Garg[3]
[1]*Research Scholar, Department of Information Technology UIET, Panjab University, Chandigarh, India*
[2]*Grade-11, Strawberry Fields High School, Chandigarh, India*
[3]*Professor, Department of Information Technology, UIET, Panjab University, Chandigarh, India*

Abstract: ***The accurate estimation of medical insurance expenses is a key requirement for risk assessment, premium determination and resource planning in health insurance systems. Prior research is usually limited by small datasets (986–2,773 records with 6–11 features) and often lacks important health, lifestyle, socioeconomic and behavioural variables, which limits its practical relevance. This study addresses these limitations. A machine learning framework is applied to a heterogeneous dataset comprising 25,000 records and 24 variables spanning demographics, medical history, behavior, economic status, and activity patterns. Missing values and categorical encoding are handled during the preprocessing of the dataset and IQRbased normalisation is applied for handling outliers and scaling. Feature selection and prediction are achieved by employing various learning paradigms such as XGBoost, Random Forest, Gradient Boosting, traditional regression models, deep learning architectures (backpropagation neural network and LSTM) and Conditional Gaussian Bayesian Networks. Model interpret ability is enhanced through explainable AI techniques. We evaluate the performance using R^2 Score, MAE and RMSE. Comparative analysis shows that larger and more diverse datasets, combined with advanced machine learning techniques, lead to more accurate and realistic cost predictions that provide practical value to insurers and policy makers. This initiative is consistent with SDG 3 (Good health and well-being) and SDG 10 (Reduced inequalities) by facilitating data-driven equitable pricing of health insurance and enhanced access to healthcare funding.***

Keywords: Medical Insurance Cost Prediction, Machine Learning, Ensemble Models, Feature Selection, Explainable AI, Multi-Feature Dataset

I. INTRODUCTION

Estimating the cost of medical insurance is important for premium pricing, risk sharing and strategy making across the insurance value chain. But existing research suffer from tiny datasets (below 3,000 records, 6–11 characteristics) which mostly comprise of basic demographics like BMI, smoking status, dependents, and area [1][2]. These constrained feature spaces are unable to capture complex real-world aspects such lifestyle habits, socioeconomic conditions, health trends, and high-risk activities, impacting model validity, fairness and application [3][4].

Advanced machine learning methods (Random Forest, Gradient Boosting, XGBoost) [5][6][7] and interpretability tools (SHAP) [8][9] are generally tested on simplified data, raising issues regarding stability, generalisation and bias in high dimensional real-world contexts. The present study provides a scalable approach employing a bigger dataset of 25,000 records with 24 variables including demographics, clinical, behavioural, socioeconomic and risk attributes. The rich feature space can capture complicated non-linear interactions that are realistic for the dynamics of the insurance cost.

Structured pre-processing pipeline: missing values, categorical encoding, normalisation and outliers (using IQR) [10][11]. The model is evaluated using R2, MAE and RMSE [12] and SHAP is employed for interpretability [8]. The methodology highlights the significance of comparative model analysis, the effectiveness of CGBN-based feature selection [10][13], and robustness to data fluctuations, and offers valuable insights for pricing, risk segmentation, and policy formulation.

II. LITERATURE REVIEW

This section reviews prior efforts on medical insurance cost prediction. Tree-based ensemble and regression models are commonly assessed on public datasets with few characteristics [1][2]. Typical explainability methods (SHAP, ICE) give insights into relevant factors (age, BMI, smoking status) [8][9].

These approaches improve interpretability, but are constrained by datasets which do not represent the complexity of the real world. The crucial weaknesses include tiny sample sizes, limited dependent and independent variables (without socioeconomic, lifestyle, and health trend factors), and minimal outside validation [3][4]. For realistic insurance pricing, largescale datasets and sound frameworks are essential.

In recent times, research has been done to merge machine learning and explainable artificial intelligence to boost the accuracy of model predictions. Ensemble approaches such as XGBoost, Gradient Boosting, and Random Forest have been used to attain higher model accuracy, supported by SHAP-based feature analysis [1]. CGBN-ML combinations have been proposed, where probability-based feature selection improves algorithm performance by reducing irrelevant variables [10][13]. Comparative regression studies have found that Gradient Boosting performs better than other techniques in terms of MAE and RMSE [2]. SHAP has also been highlighted alongside ensemble models for personalizing insurance premiums [5], underlining the increasing importance of interpretability in healthcare applications.

Ensemble models such as XGBoost and Random Forest are promising candidates to simulate complex non-linear associations and feature interactions in structured healthcare data [6][7]. Deep learning (ANN, LSTM) is good for temporal data such as electronic health records, but requires massive datasets [16]. Bayesian Networks can also undertake principled feature selection, i.e., they can identify important predictors and correlations such that redundancy is reduced and efficiency is improved [13][11].

Evaluation typically employs MAE (robust to outliers) and RMSE (penalizes large errors heavily) [12], alongside growing emphasis on fairness and bias analysis for ethical insurance pricing and healthcare decisions [4].

Despite these progresses, research are constrained by low feature diversity and short dataset size, which hinder generalisation in the real-world with high variability and complexity. Hence, a complete framework is needed that integrates largescale multi-dimensional datasets, advanced machine learning (ML), probabilistic feature selection, and explainable AI to improve both predictive performance and practical applicability.

III. SYSTEM MODEL AND PROBLEM FORMULATION

A. Dataset Description

This study uses a Kaggle dataset from Karthikeyanrajuz [15]. It contains 25,000 anonymised patient records including demographic, lifestyle, medical history and insurance attributes. This is a great foundation for healthcare analytics and is well suited to predicting medical insurance costs and risk profiling.

To perform a structured analysis, the characteristics of the datasets are categorised according to their functional relevance:

1) Demographic Details: Patient ID, Age, Gender, Location, and Occupation.
2) Lifestyle Factors: Alcohol Consumption, Smoking Status, Average Daily Steps, Exercise Regimen, and Weight Change in the Last Year.
3) Medical History: Heart Disease History, Other Major Illnesses, Cholesterol Levels, Average Glucose Level, Body Mass Index (BMI), Body Fat Percentage, Number of Doctor Visits, and Regular Checkups.
4) Insurance Details: Years with Current Provider, Last Admission Year, Coverage by Other Insurance Providers, and Participation in Adventure Sports.
5) Outcome variable: Patient insurance cost, denoting the aggregate medical insurance expenditure attributed to each individual.

The dataset is broad and deep enough to predict complex health, lifestyle and financial risk correlations making the ML framework robust and usable in real world scenarios.

B. Data Preprocessing

Robust prediction models in healthcare necessitate data preprocessing to tackle issues like inconsistencies, missing values and heterogeneous characteristics [10][11]. In this study, a structured framework is used to clean, transform and encode the dataset for high-quality data for machine learning algorithms.

1) Data Cleaning

- Redundant patient data was discarded to avoid bias fromduplicate patient records.
- Columns that did not provide information were deleted,leaving a data set with 25,000 records and 24 attributes.

2) Missing Value Imputation:

- Numerical features: Median imputation was used to control for skewness.
- Categorical features: Mode imputation preserved classdistribution.
- Special case: patient year last admitted missing values were set to –1 to indicate no prior admissions.

3) Handling Infinite Values: Infinite values (positive/negative) were replaced with NaN and imputation was done using the median of the feature. All undefined values were excluded.

4) Outlier Detection and Treatment: The Interquartile Range (IQR) technique was used to identify outliers [11]. Instead of deleting outliers, they were capped at limits, such that all records were kept and their influence reduced. The desired variable (patient insurance cost) is removed and the changes in the real costs are retained. IQR is defined as:

$$IQR = Q3 - Q1 \quad (1)$$

fig.1. shows the feature distributions without IQR treatment (e.g. outliers in BMI, glucose, steps), while fig.2. shows the cleaner and more stable distributions after treatment, validating the data quality improvement.

5) Feature Categorization and Encoding:

- Numerical features (age, BMI, body fat, daily steps,insurance years) retained as continuous.
- Multi-class categorical (occupation, region) encoded viaOne-Hot Encoding.
- Binary features (smoking, disease history) mapped to 0/1.
- Ordinal features (alcohol consumption, exercise, cholesterol) used custom rank mappings.

6) Feature Scaling: z-score normalisation (StandardScaler) was used to scale numerical variables to normalise to similar ranges and avoid bias to high range features [11].

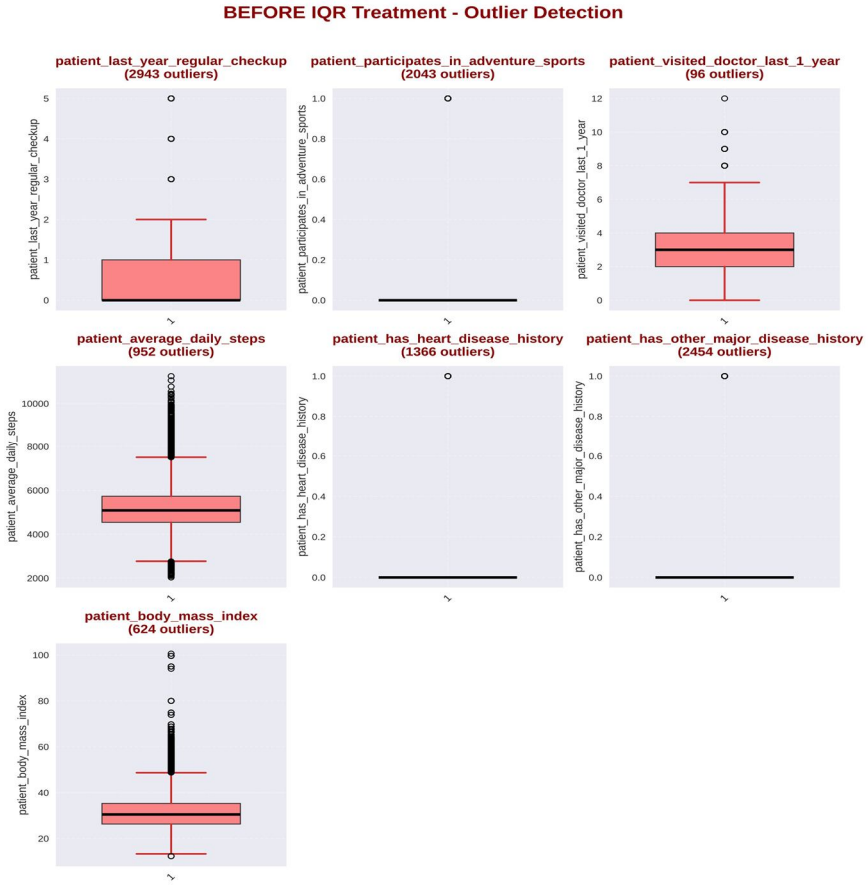


Fig.1. Boxplots of numerical features before IQR treatment showing outliers

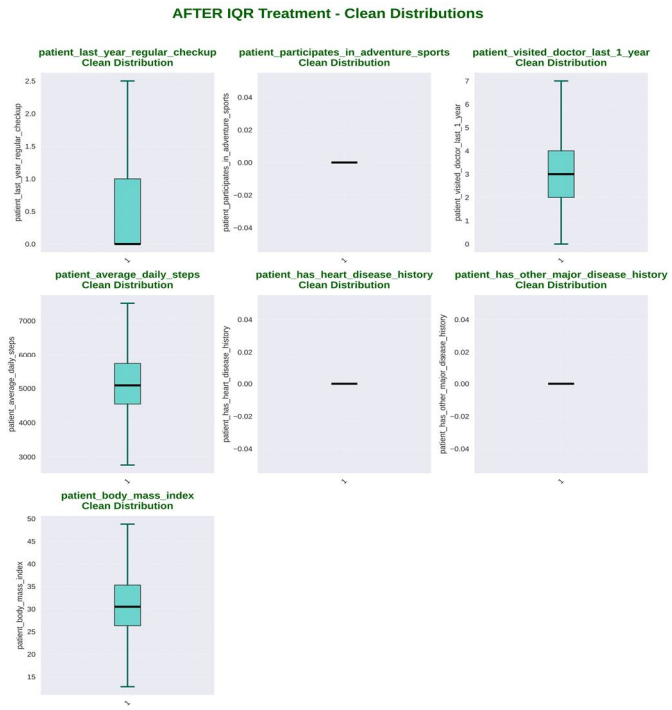


Figure 2. Boxplots after IQR treatment showing cleaned distributions

7) *Preprocessing Pipeline:* To achieve reproducibility, modularity and scalability we employed a scikit-learn pipeline (ColumnTransformer + Pipeline) by applying imputation, encoding and scaling consistently to training and testing data [11]. The dataset was changed to be fully quantitative, tidy and standardised for several machine learning techniques.

C. Techniques Used

This work employs a variety of machine learning models and analytical techniques to enhance the predictive performance and interpretability of medical insurance expense estimation:

- Parametric model Linear Regression Model Assumes a linear relationship Parameter estimation via minimising squared errors.
- Ridge Regression: Linear regression, with L2 regularisation to reduce overfitting and deal with multicollinearity.
- Decision Tree Regression: Feature space partitioning of model, predicting the mean value of leaf nodes, so as to capture non-linearities.
- Random Forest Regression: An ensemble of decision trees that uses bootstrap sampling and random feature selection to improve accuracy and reduce variance [7].
- Gradient Boosting Regression: Sequential ensemble fitting residuals to optimize loss functions iteratively.
- XGBoost: Efficient gradient boosting with regularisation, second order optimisation and fast tree building [6].
- Support Vector Regression (SVR): Fits data within ϵinsensitive margin using kernels for nonlinear relationships.
- K-Nearest Neighbours (KNN) Regression: Instance based method that averages K-nearest neighbours by the distance metric.
- Backpropagation Neural Network (BPNet): Complex nonlinear relationships trained by backpropagation in feedforward network.
- Long Short-Term Memory (LSTM): Recurrent neural network with gated memory for long-term sequential dependencies [16].
- Conditional Gaussian Bayesian Network (CGBN): Probabilistic graphical model for feature selection with HC, Tabu, PC, GS, MMHC, and rsmax2 algorithms to remove redundant features [10][13].
- Explainable AI (XAI): Techniques to improve ML model transparency, relevant for healthcare applications [8][9].
- SHAP for Feature Attribution: A model-agnostic method for attributing SHAP values to feature contributions to the predictions [8].

D. Performance Evaluation Metrics

Standard metrics measuring the difference between actual and anticipated values were used to evaluate and compare the regression models, and to assess their accuracy and generality.

Mean Absolute Error (MAE) measures average error magnitude:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \quad (2)$$

Root Mean Squared Error (RMSE) squares then averages errors, emphasizing larger deviations:

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

R^2 Score indicates explained variance:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (4)$$

IV. PROPOSED METHODOLOGY

As demonstrated in fig. 3., we present a comparative machine learning framework for medical insurance cost prediction. The framework is based on two approaches namely: (i) straight regression modelling and (ii) hybrid approach of Conditional Gaussian Bayesian Networks (CGBN) and regression models [10][13]. Both undergo a common pipeline of data pre-processing, model training, evaluation and interpretation.

A. Framework Overview

This methodology starts with Kaggle dataset collecting [15] followed by data preprocessing which involves data cleaning, addressing missing values, encoding categorical variables, feature scaling, and outlier treatment using Interquartile Range (IQR) method [11]. The cleaned and preprocessed data are supplied to the two methodological frameworks for comparison.

B. Regression Based Modeling

The first option is direct application of regression models on the preprocessed dataset. These models are Linear Regression, Ridge Regression, Decision Tree, Random Forest, Gradient Boosting, XGBoost, Support Vector Regression (SVR), KNearest Neighbours (KNN), Backpropagation Neural Network (BPNet) and Long Short-Term Memory (LSTM) [6][7][16]. The training dataset is used to train each model and both training and testing datasets are used for evaluation to check the performance and capabilities of generalisation.

C. Hybrid Modeling with CGBN and Regression Based Modeling

The second approach presents a hybrid framework using Conditional Gaussian Bayesian Networks (CGBN) for feature selection [10][13]. CGBN learns the probabilistic relationships among the variables and eliminates the redundant and irrelevant features. The optimal feature set is then utilised as an input to the same set of regression models as in the first technique, in order to allow a direct comparison between the original and optimised feature spaces and to investigate the effect of probabilistic feature selection over the model performance.

D. Model Training and Evaluation

Both methodologies are evaluated on the dataset using standard protocols that partition it into training and testing subsets. The models are trained with the training data and evaluated with novel test data. To evaluate the performance, three standard measures for evaluation of regression are used [12]:

- Mean Absolute Error (MAE)
- Root Mean Squared Error (RMSE)
- Coefficient of Determination (R^2)

These metrics provide a comprehensive evaluation of prediction accuracy, error magnitude, and model generalization.

E. Analysis of Top Models

The top three models from each approach are selected based on evaluation metrics for further analysis using:

- Learning curves to analyse training and validation performance and to detect overfitting or underfitting.
- Boxplots to visualise feature distribution and the impact of preprocessing methods like outlier treatment.
- Heatmaps for investigating correlations of features and relationships between variables.
- SHAP analysis to explain the model predictions by quantifying the feature importance at the global and local levels [8][9].

This analysis enhances model interpretability and provides deeper insights into the factors influencing medical insurance cost prediction.

F. Final Prediction

Comparative evaluation and analysis selects the ideal model, and obtains the final forecast of medical insurance costs. Our architecture is highly predictive and interpretable, which makes it suitable for real-world healthcare and insurance applications [1][3][5].

V. RESULTS AND DISCUSSION

A. Normal Regression-Based Modeling

This section utilises the machine learning models outlined previously for the prediction task and assesses performance through Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and R2 score, which measures prediction error and the variance explained by the model [12]. The ensemble models have better predictive performance as shown in Table I. Gradient Boosting performs very well with the highest test R2 score of (95.64%) and relatively low RMSE and MAE values. This is due to the sequential learning process which can model complicated nonlinear interactions and is less likely to have prediction errors [1][6].

The second best model is XGBoost with the test R2 of 95.55%. It has built-in regularisation and good optimisation although the levels of error are a bit high [6]. This means that Gradient Boosting achieves a better bias-variance trade-off on this dataset. Both models do well, while Gradient Boosting gives a better balance of accuracy and generality.

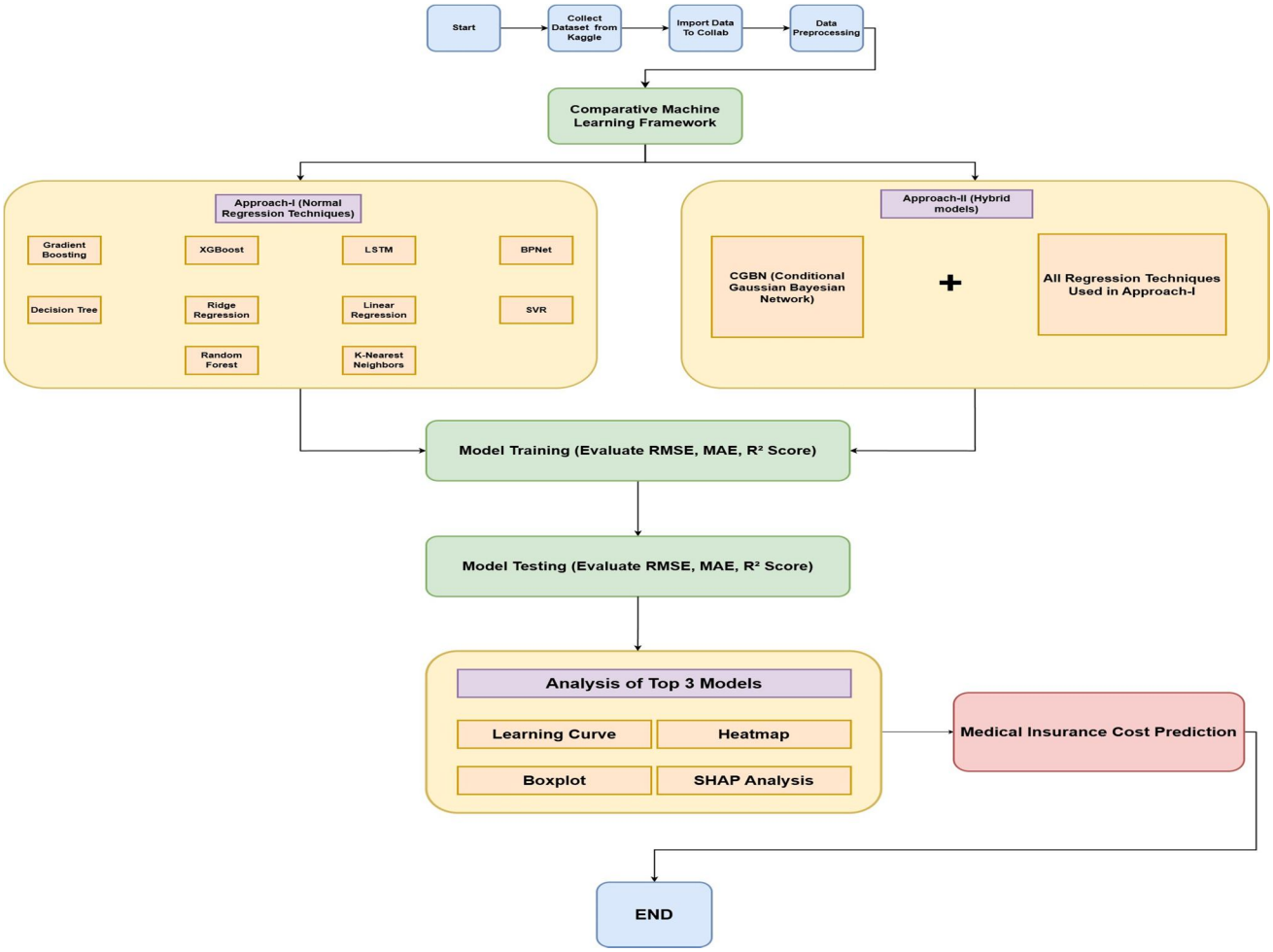


Fig.3. Comparative Machine Learning Framework for Medical Insurance Cost Prediction

TABLE I: Performance Comparison of Regression-Based Modeling

Model	TR RMSE	TE RMSE	TR MAE	TE MAE	TR R^2 (%)	TE R^2 (%)
Gradient Boosting	2436.39	2981.65	1954.17	2379.19	97.11	95.64
XGBoost	2117.94	3011.19	1677.31	2398.20	97.82	95.55
Random Forest	1133.11	3021.38	898.80	2401.32	99.38	95.52
LSTM	3010.02	3103.71	2433.90	2516.96	95.59	95.27
BPNet	2739.67	3224.88	2170.44	2599.97	96.35	94.90
Decision Tree	2629.83	3247.75	2049.17	2532.19	96.63	94.82
Linear Regression	3354.64	3355.75	2709.30	2716.83	94.52	94.47
Ridge Regression	3354.64	3355.74	2709.29	2716.81	94.52	94.47
SVR	4726.86	4798.06	3589.05	3656.50	89.13	88.70
KNN	0000.00	6531.88	0000.00	5202.75	100.00	79.06

We have chosen the three leading models (Gradient Boosting, XGBoost, Random Forest) for subsequent study (refer to Table I). fig. 4. evaluates these models based on RMSE, MAE, and R^2 across training and testing datasets. Gradient Boosting exhibits the most balanced performance, characterised by low error values and consistently elevated R^2 values across both datasets, signifying effective generalisation. XGBoost exhibits comparable performance with marginally greater errors, however maintains a consistent R^2, signifying robustness in structured data [6]. The Random Forest has strong performance on the training set; however, there is a noticeable rise in RMSE and MAE on the test set, suggesting modest overfitting, albeit maintaining a high R^2 value [7].
The learning curves of the three best models are presented

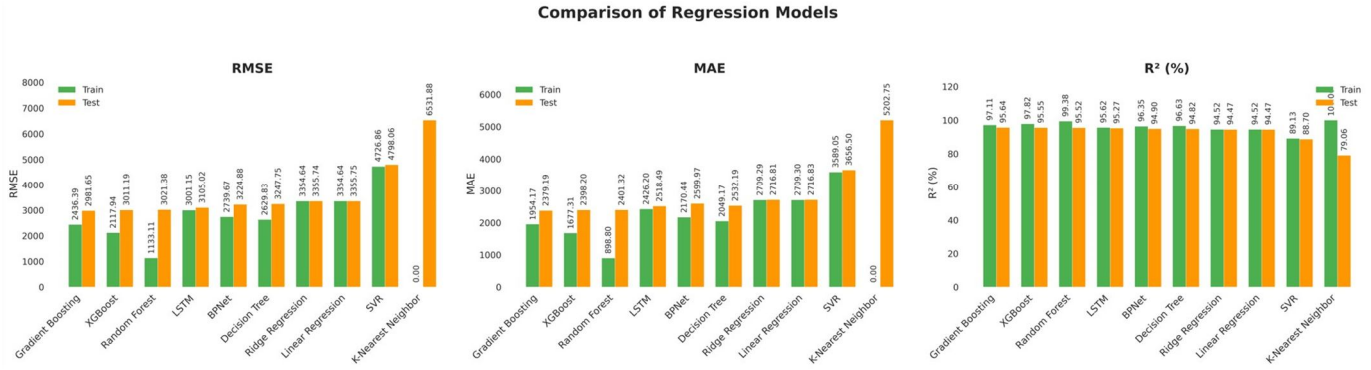


Figure 4. Comparative performance of the best three models

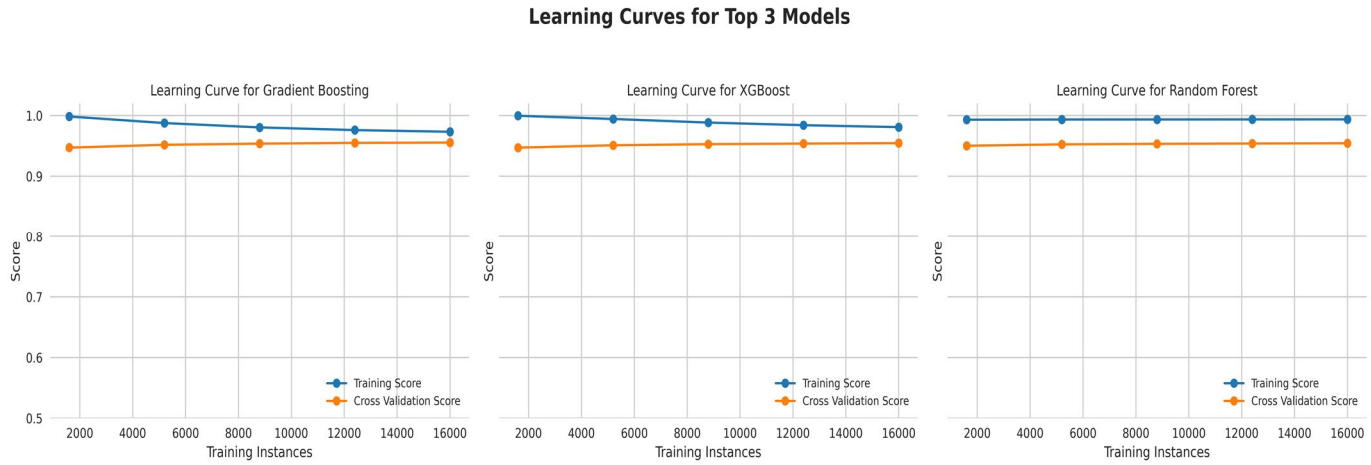


Figure 5. Learning curves of the best three models

in fig. 5. In both Gradient Boosting and XGBoost the training and validation curves tend to converge extremely closely with increase of dataset size, indicating good generalisation with very little overfitting and favourable bias-variance trade-off. The random forest has a decent training performance and the training and validation curve are further apart, which confirms that there is a minor overfitting, but the validation performance is still robust [6][7].

fig. 6 presents the SHAP-based feature importance for Gradient Boosting, XGBoost and Random Forest. Among these, *patient weight* is the most important feature with the biggest range of SHAP value and positive correlation with estimated costs. *patient year last admitted*, insurance coverage and regular checkups have moderate effect, whereas age, BMI, glucose level and daily steps have minimal effect.

The stable rankings suggest that the acquired patterns are resilient. In conclusion, SHAP analysis indicates that lifestyle and healthcare use are the primary determinants of insurance costs, whereas other variables exert minimal influence on prediction variability [8][9].

B. Hybrid Modeling with CGBN and Normal RegressionBased Modeling

Conditional Gaussian Bayesian Networks (CGBN) learn the probabilistic dependencies to identify and remove redundant features [10][13]. This process reduces dimensionality and

TABLE II: Performance Comparison of CGBN-Based Hybrid Models

Model	TR RMSE	TE RMSE	TR MAE	TE MAE	TR R^2 (%)	TE R^2 (%)
CGBN + Random Forest	6471.74	6734.11	5177.22	5395.08	79.62	77.74
CGBN + Gradient Boosting	6818.68	6735.34	5454.53	5406.86	77.38	77.73
CGBN + XGBoost	6810.67	6735.82	5448.00	5404.99	77.43	77.73
CGBN + BPNet	6901.21	6832.47	5484.12	5446.52	76.83	77.09
CGBN + Decision Tree	6329.99	6899.58	5066.60	5525.64	80.50	76.63
CGBN + Linear Regression	7027.07	6960.48	5592.70	5561.57	75.97	76.22
CGBN + Ridge Regression	7027.38	6960.78	5592.95	5561.82	75.97	76.22
CGBN + SVR	8113.35	8076.39	6368.83	6379.44	67.97	67.98
CGBN + KNN	8298.62	8774.49	6616.57	7027.96	66.49	62.21
CGBN + LSTM	13098.55	12981.78	10736.72	10635.58	16.50	17.30

retains informative predictors, thus improving the model efficiency, interpretability and predictive performance.

As shown in Table II, the best performance of the hybrid models is achieved by CGBN + Random Forest, which yields the lowest test RMSE (6734.11) and the highest R^2 (77.74%). The reason behind this great performance is ensemble based modelling of non-linear relationship [7]. Boosting based variations such as CGBN + Gradient Boosting and XGBoost are competitive as well as they iteratively reduce mistakes [6]. However, models like SVR, KNN, and LSTM are not as applicable to the limited feature space. The ensemble methods are still the best again since linear models are not able to capture complicated patterns.

Fig. 7. shows the distribution of test R^2 for the hybrid models. Ensemble approaches such as Random Forest, Gradient Boosting and XGBoost attain the greatest performance (approx. 77.7%) indicating the good performance of the feature selection. On the contrary, the performance of SVR, KNN and LSTM is poorer, which means their less applicability to this feature subspace.

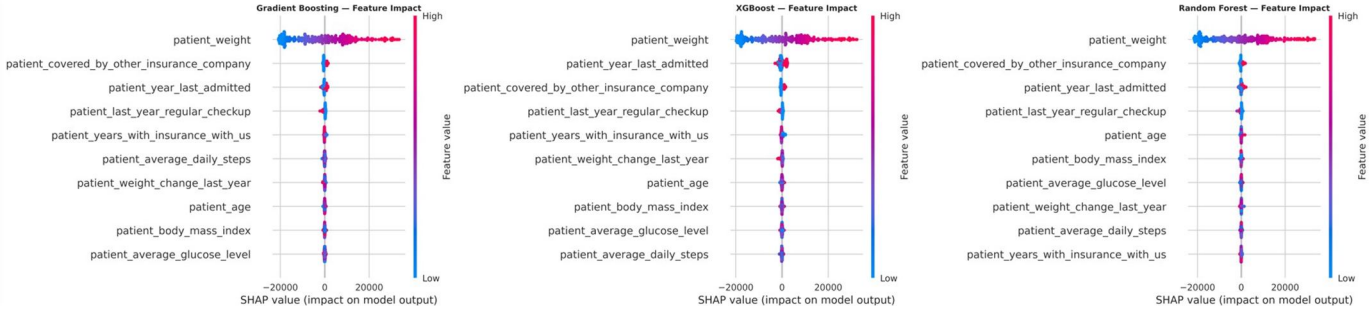


Figure 6. SHAP-based feature importance analysis for the best three models

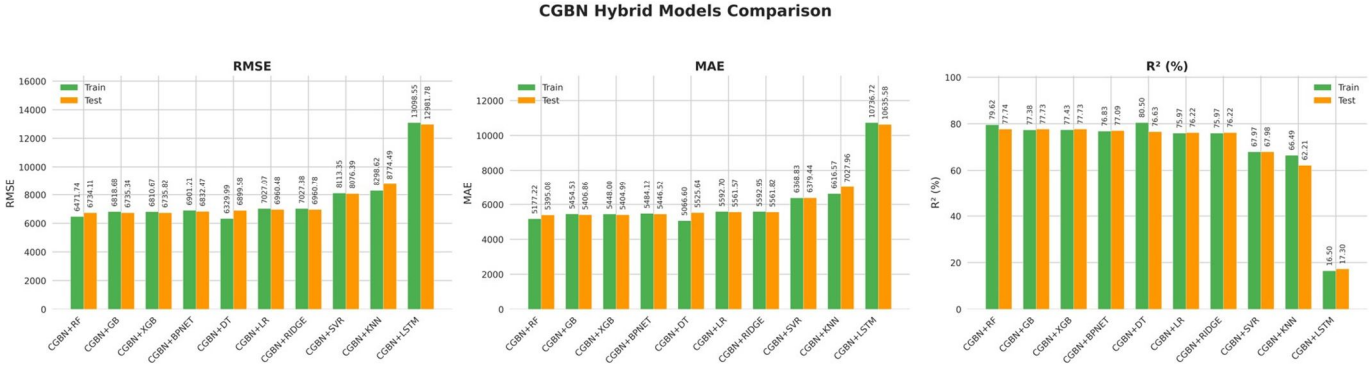


Figure 7. Comparative performance of the best three CGBN-based hybrid models

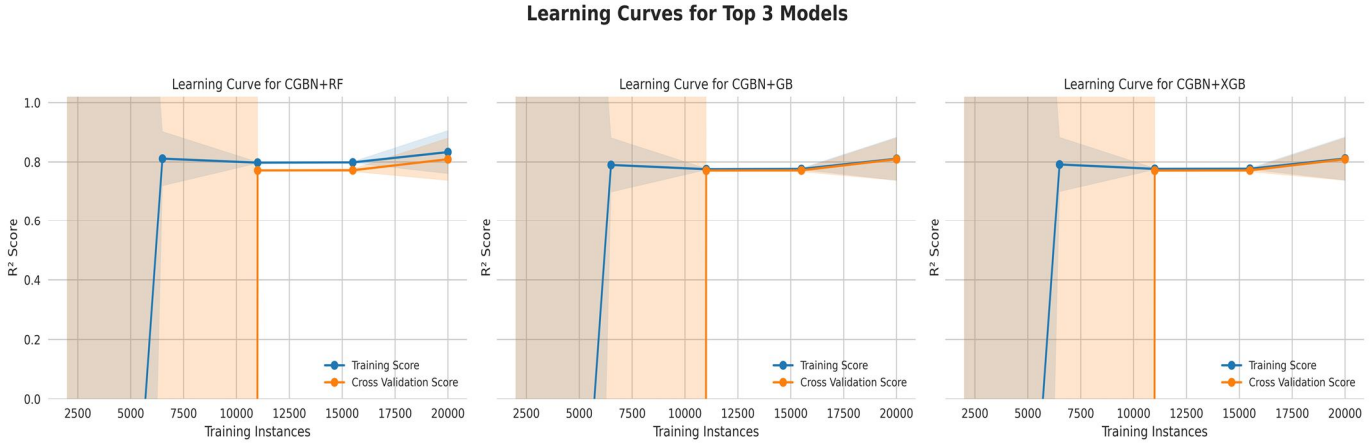


Figure 8. Learning curves for best three hybrid models

Fig. 8. shows the training and validation R^2 scores for the three best hybrid models. The tight agreement of these curves suggests that the generalisation is powerful and not much over fitting occurs. Random Forest has a bit better training score, but the Gradient Boosting and XGBoost have more balanced learning behaviour in their curves, confirming stable performance in all top models [6][7].

1) *SHAP Analysis:* fig. 9. shows the SHAP analysis of the best hybrid models [8][9]. The most dominant attribute is *patient weight* that has a positive impact on the insurance cost estimates. Hospital admission, insurance coverage and health check-ups have modest influence, BMI, age, glucose level and daily steps have minor impact. The reduced number of highimpact features relative to the first strategy validates CGBN's capability to filter out redundant variables for improved interpretability and consistent feature relevance across models [10][13].

C. Comparative Analysis with Existing Methodologies

The performance of the proposed CGBN-based hybrid framework is evaluated against existing studies [1][10][2][5], as detailed in

figure 9. SHAP-based feature importance analysis for the best three models

TableIII.

TABLE III: Comparison Of The Proposed Technique With Existing Approaches

Ref.	Dataset	Regression Techniques	Feat.	Accuracy
[1]	Kaggle ”Medical Insurance”; 986 records	XGBoost, Gradient Boosting, RF	11	86.47%
[10]	Two Kaggle datasets: 986 / 1,338 records	CGBN +Regressors (RF, SVR, BP NN, LR, LSTM)	10 / 6	92.9% / 79%
[2]	Insurancedataset: 2,773 records	Linear, Ridge, DT, RF, KNN, SVR, XGBoost, GB	7	86.79%
[5]	Primarysurveydataset (n=145) (India)	MultipleLR,RF, Gradient Boosting	23	61.00%
Prop.	Kaggle dataset: 25,000 records	Normal Regression Models & CGBN Hybrid models	24	95.64% & 77.74%

VI. CONCLUSION AND FUTURE SCOPE

The results indicate that ensemble learning approaches, such as Random Forest, Gradient Boosting and XGBoost, perform better than standard regression models in terms of accuracy and generalization [1] [6] [7].

We propose an interpretable machine learning framework for medical insurance cost prediction on a large heterogeneous dataset of 25,000 records. To offer high quality inputs for Normal Regression-Based Modelling and a hybrid Conditional Gaussian Bayesian Network (CGBN)-based framework, we developed a single pre-processing pipeline including imputation, encoding, scaling and IQR-based outlier treatment.

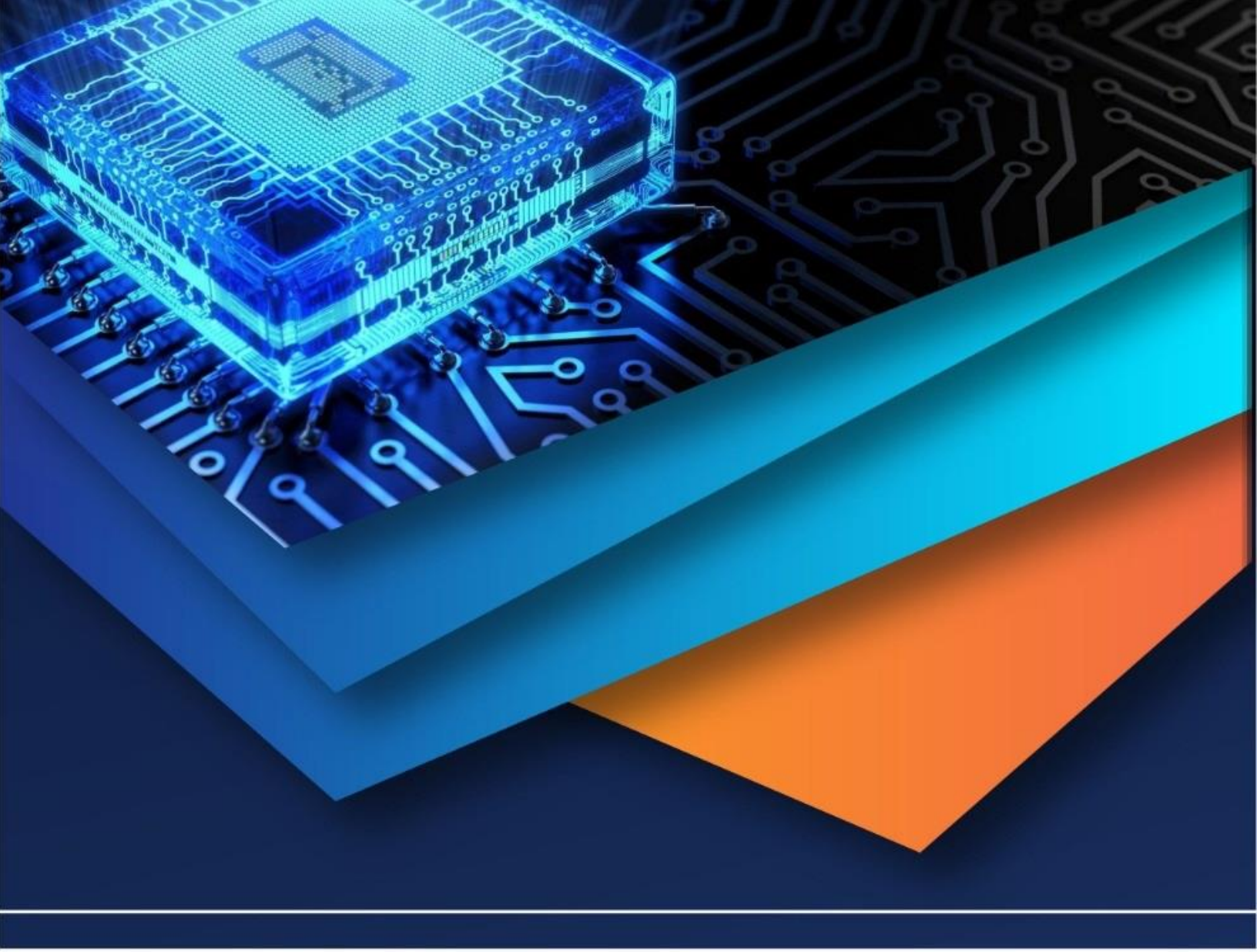
In the experimental assessment, the ensemble models outperformed the individual models in terms of R^2 Score, MAE, and RMSE, obtaining a maximum R^2 Score of about 95.64%. Moreover, we established the higher significance of healthcare use and lifestyle-related variables compared to demographic features in predicting insurance cost using SHAP analysis.

The hybrid CGBN approach showed how probabilistic feature selection improves interpretability and dimensionality reduction. The prediction accuracy fell in the limited feature space (test $R^2 \approx 77.74\%$) but remained steady in proportion to the significance of features and improved transparency, indicating the trade-off between accuracy and explainability.

The overall results suggest that the scalable and policy relevant medical insurance cost prediction is doable with the combination of robust pre-processing, feature dense datasets, ensemble learning and explainable AI. Future work will formalize a composite Lifestyle and Behavioural Health Risk Index (LBHRI), integrate longitudinal wearable data streams, and incorporate fairness-aware learning to enhance ethical, personalized, and dynamic risk prediction.

REFERENCES

[1] U. Orji and E. Ukwandu, "Machine learning for an explainable cost prediction of medical insurance," Machine Learning with Applications, vol. 15, Art. no. 100516, 2024. doi: 10.1016/j.mlwa.2023.100516.
[2] M. M. Billa and T. Nagpal, "Medical insurance price prediction using machine learning," Journal of Electrical Systems, vol. 20, no. 7, pp. 2270–2279, 2024.
[3] A. Rajkomar, J. Dean, and I. Kohane, "Machine learning in medicine," New England Journal of Medicine, vol. 380, no. 14, pp. 1347–1358, 2019. doi: 10.1056/NEJMra1814259.
[4] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," ACM Computing Surveys, vol. 54, no. 6, pp. 1–35, 2021. doi: 10.1145/3457607.
[5] D. 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 prediction in Sierra Leone," International Journal of Medical Informatics, vol. 211, Art. no. 106313, 2026. doi: 10.1016/j.ijmedinf.2026.106313.
[6] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining, 2016, pp. 785–794. doi: 10.1145/2939672.2939785.
[7] R. Genuer and J.-M. Poggi, "Random forests," in Random Forests with R. Cham, Switzerland: Springer, 2020, pp. 33–55. doi: 10.1007/978-3030-56485-8 3.
[8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in Advances in Neural Information Processing Systems, vol. 30, 2017.
[9] P. Biecek and T. Burzykowski, Explanatory Model Analysis: Explore, Explain, and Examine Predictive Models. Boca Raton, FL, USA: Chapman and Hall/CRC, 2021.
[10] S. Zou, C. Chu, N. Shen, and J. Ren, "Healthcare cost prediction based on hybrid machine learning algorithms," Mathematics, vol. 11, no. 23, Art. no. 4778, 2023. doi: 10.3390/math11234778.
[11] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of Machine Learning Research, vol. 3, pp. 1157– 1182, 2003.
[12] C. J. Willmott and K. Matsuura, "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance," Climate Research, vol. 30, no. 1, pp. 79–82, 2005.
[13] D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and Techniques. Cambridge, MA, USA: MIT Press, 2009.
[14] M. Kapse, V. Sharma, R. Vidhale, and V. Vellanki, "Customization of health insurance premiums using machine learning and explainable AI," International Journal of Information Management Data Insights, vol. 5, no. 1, Art. no. 100328, 2025. doi: 10.1016/j.jjimei.2025.100328.
[15] karthikeyanrajuz, "Medical Insurance Prediction Dataset," Kaggle, [Online]. Available: https://www.kaggle.com/datasets/karthikeyanrajuz/med ical-insurance-prediction-dataset. [Accessed: n.d.].
[16] B. Shickel, P. J. Tighe, A. Bihorac, and P. Rashidi, "Deep EHR: A survey of recent advances in deep learning techniques for electronic health record (EHR) analysis," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1589–1604, 2018. doi: 10.1109/JBHI.2017.2767063.
[17] B. Wiley and V. Fuster, "The concept of the polypill in the prevention of cardiovascular disease," Annals of Global Health, vol. 80, no. 1, pp. 24–34, 2014. doi: 10.1016/j.aogh.2013.12.008.
[18] World Health Organization, Global Health Risks: Mortality and Burden of Disease Attributable to Selected Major Risks. Geneva, Switzerland: WHO Press, 2009.
[19] Y. Mahwati et al., "Development of machine learning models to predict health insurance claim costs among older Indonesians: A retrospective predictive modeling study," Journal of Preventive Medicine and Public Health, vol. 59, no. 2, pp. 132–142, 2026. doi: 10.3961/jpmph.25.350.
[20] D. K. Kadali et al., "Analysis and prediction of health insurance cost using machine learning approaches," in Proc. Int. Conf. Computational Innovations and Emerging Trends *(ICCIET)*, 2024, pp. 569–577. doi: 10.2991/978-94-6463-471-6 55.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY