Bangladesh Army University of Science and Technology (BAUST)

Saidpur, Nilphamari, Bangladesh

Department of Computer Science and Engineering

Health Insurance Cost Prediction Using Machine Learning

A Comprehensive Comparative Analysis

Submitted By:

Masuma Sadia Sagota
220201088

Supervised By:

Engr. Rohul Amin
Lecturer

Nadim Reza
Lecturer

November 25, 2025

Health Insurance Cost Prediction Using Machine Learning: A Comprehensive Comparative Analysis

Masuma Sadia Sagota
Department of Computer Science and Engineering
Bangladesh Army University of Science and Technology
Saidpur, Nilphamari, Bangladesh
masumasadia88@gmail.com

***Abstract*—Accurately forecasting individual health insurance charges is a critical regression problem for actuarial science, premium pricing, and healthcare risk management. This study develops a comparative machine learning framework to model the non-linear dependencies between demographic and lifestyle factors—including age, BMI, and smoking status—and medical expenditures. Using the publicly available insurance.csv dataset containing 1,338 records, the methodology integrates a complete preprocessing pipeline with domain-driven feature engineering, most notably the Smoker–BMI interaction feature, which captures the amplified risk associated with smoking among high-BMI individuals. Seven regression models were evaluated and optimized using GridSearchCV: Linear Regression, Ridge Regression, Decision Tree, Random Forest, XGBoost, and Gradient Boosting. The final optimized model, the Optimized Random Forest Regressor, achieved the best performance with an R^2 score of 0.8703, a Mean Absolute Error (MAE) of $2,535.04, and a Root Mean Squared Error (RMSE) of $4,481.58, significantly outperforming all other models and surpassing the best reported academic score (0.8538) by 1.65%, demonstrating the superior predictive power of combining ensemble learning with targeted, domain-specific feature engineering.**

***Index Terms*—Health Insurance Cost Prediction, Regression Analysis, Random Forest, Gradient Boosting, Feature Engineering, Ensemble Learning, Machine Learning, Healthcare Analytics, Premium Pricing**

I. Introduction

A. Background and Motivation

Health insurance cost prediction enables providers to determine fair premium rates, reduce financial risk, and design informed policy interventions. Medical charges depend on multiple interacting features—age, sex, BMI, number of dependents, smoking behavior, and geographic region—creating a highly non-linear prediction landscape that traditional statistical models often fail to capture effectively.

Accurate prediction of health insurance costs is essential for insurance companies to maintain profitability while offering competitive rates. For individuals, transparent cost prediction enables informed decision-making regarding coverage options. For policymakers, understanding cost drivers facilitates the design of targeted public health interventions.

Despite significant research in this domain, many existing approaches suffer from inadequate preprocessing, lack of feature engineering, and failure to properly handle the non-linear interactions between predictive variables. This study addresses these limitations through a comprehensive machine learning framework that emphasizes robust preprocessing, domain-driven feature engineering, and systematic model comparison.

B. Research Objectives

The primary objectives of this study are:

Implement a complete preprocessing workflow including proper handling of missing values, encoding of categorical variables, and standardization of numerical features.

Engineer domain-specific interaction features to capture high-risk combinations, particularly the Smoker–BMI interaction.

Train and compare seven regression models under identical conditions to establish fair performance benchmarks.

Optimize the best-performing model using GridSearchCV with cross-validation.

Achieve state-of-the-art prediction accuracy that surpasses published research on the same dataset.

Establish a reliable, deployment-ready solution for insurance cost prediction.

C. Major Contributions

This research makes the following key contributions:

State-of-the-Art Performance: Achieved $R^2 = 0.8703$, surpassing the best reported academic score (0.8538) on the same dataset by 1.65%.

Domain-Driven Feature Engineering: Introduced the Smoker_BMI interaction feature, capturing the multiplicative cost effect when high BMI and smoking co-occur.

Optimized Ensemble Architecture: Applied GridSearchCV with 5-fold cross-validation to tune ensemble models for maximum generalization.

Robust Preprocessing Pipeline: Implemented a reproducible pipeline including proper missing value imputation, categorical encoding, and feature scaling.

Comprehensive Model Benchmarking: Seven algorithms compared using consistent metrics (R^2, MAE, RMSE) under the same data split.

D. Paper Organization

The remainder of this paper is organized as follows: Section II reviews related work in health insurance cost prediction. Section III describes the dataset characteristics and preprocessing pipeline. Section IV details the methodology, including feature engineering and model architectures. Section V presents experimental setup and results. Section VI discusses findings and practical implications. Section VII concludes the paper and suggests future research directions.

II. Related Work

A. Literature Review

Health insurance cost prediction has emerged as a critical application of machine learning in actuarial science and healthcare analytics. The ability to accurately predict individual medical expenses enables insurance companies to optimize premium pricing strategies, maintain financial sustainability, and design risk-mitigation policies. This section reviews significant contributions to health insurance cost prediction, highlighting methodologies, datasets, and performance achievements while identifying research gaps that motivate the present study.

Kumar et al. [1] applied Simple Linear Regression and Random Forest models to predict health insurance costs using the insurance.csv dataset. Their Random Forest model achieved $R^2 = 0.8538$, representing the best reported performance prior to this study. However, their approach suffered from inadequate preprocessing—specifically, they did not apply standardization to numerical features, which can significantly impact model performance, particularly for distance-based and gradient-based algorithms. The absence of feature engineering meant their model relied solely on raw demographic attributes without capturing critical interactions such as the combined effect of smoking and obesity.

Aishwarya et al. [2] developed a health insurance cost prediction application using Random Forest on the insurance dataset, achieving $R^2 = 0.83$. While their application provided practical value through a user-friendly interface, the study exhibited minimal feature engineering and lacked systematic comparison with alternative algorithms. The absence of proper cross-validation and hyperparameter tuning limited generalization performance. Their work focused primarily on application development rather than methodological rigor or predictive accuracy optimization.

Sharma et al. [3] implemented Linear Regression in R for medical insurance cost prediction using the insurance dataset, achieving adjusted $R^2 = 0.7489$. Their approach did not include standardization or proper encoding of categorical variables, resulting in suboptimal performance. The study highlighted the limitations of linear models when applied to datasets with complex non-linear interactions but did not explore ensemble methods or interaction features that could have improved performance. Their analysis provided valuable insights into baseline model behavior but lacked the sophistication needed for production deployment.

Paul et al. [4] explored K-Nearest Neighbors (KNN) Regression in R for health insurance cost prediction, achieving $R^2 = 0.5521$ on the insurance dataset. This poor performance demonstrates KNN's high sensitivity to scaling issues in high-dimensional data. The study did not implement proper feature scaling, which is critical for distance-based algorithms like KNN. Additionally, no hyperparameter optimization was performed for the number of neighbors (k), further limiting model performance. The weak results underscore the importance of proper preprocessing and algorithm selection for regression tasks.

B. Feature Engineering in Healthcare Prediction

Hunter et al. [5] investigated the interaction between smoking and obesity in healthcare costs, demonstrating that the combined effect significantly exceeds the sum of individual effects. Their domain analysis motivated our inclusion of the Smoker_BMI interaction feature, which proved to be a critical predictor in our model.

C. Ensemble Methods in Regression

Ensemble methods, particularly Random Forest and Gradient Boosting, have consistently demonstrated superior performance in regression tasks across various domains. Random Forest's bagging approach reduces variance by averaging predictions from multiple decorrelated decision trees. Gradient Boosting's sequential learning mechanism allows each tree to correct errors from previous trees, enabling the model to capture complex patterns.

D. Summary of Related Work

Table I presents a comprehensive comparison of prior research in health insurance cost prediction, summarizing the R^2 performance, methodologies, and key findings from each study.

TABLE I
Summary of Related Work

Paper	R^2 Score	Methodology
Kumar et al. [1]	0.8538	Random Forest; No standardization; Insurance.csv dataset (1,338 records)
Aishwarya et al. [2]	0.83	Random Forest; Web application development; Minimal feature engineering
Sharma et al. [3]	0.7489	Linear Regression in R; No categorical encoding; No standardization
Paul et al. [4]	0.5521	KNN Regression in R; No feature scaling; No hyperparameter tuning
This Study	**0.8703**	**7 algorithms; Feature engineering (Smoker_BMI); GridSearchCV optimization; Optimized Random Forest**

E. Research Gaps

While existing research has demonstrated the potential of machine learning for health insurance cost prediction, several gaps remain:
Limited systematic comparison of multiple regression algorithms under identical preprocessing conditions.
Insufficient feature engineering to capture domain-specific interactions such as smoking-obesity effects.
Lack of proper hyperparameter optimization in most studies, leading to suboptimal model performance.
Inconsistent evaluation methodologies making cross-study comparisons difficult and unreliable.
Absence of reproducible preprocessing pipelines that could be adapted for production deployment.
This study addresses these gaps by providing a comprehensive evaluation framework with robust preprocessing, domain-driven feature engineering, and systematic model optimization.

III. Dataset and Preprocessing

A. Dataset Description

The dataset used in this study is the publicly available insurance.csv containing 1,338 employee records with 7 features and 1 target variable (charges).

Features include:
age: Age of the primary beneficiary (continuous, 18–64 years).
sex: Gender of the beneficiary (categorical: male, female).
bmi: Body Mass Index (continuous, 15.96–53.13).
children: Number of dependents covered by insurance (discrete, 0–5).
smoker: Smoking status (binary: yes, no).
region: Beneficiary's residential region (categorical: southwest, southeast, northwest, northeast).
charges: Individual medical costs billed by insurance (continuous, target variable).

Descriptive Statistics:
Age: mean = 39.21, std = 14.05, range = [18, 64].
BMI: mean = 30.66, std = 6.10, range = [15.96, 53.13].
Children: mean = 1.09, std = 1.21, range = [0, 5].
Charges: mean = \$13,270.42, std = \$12,110.01, range = [\$1,121.87, \$63,770.43].
Sex distribution: female (662), male (676).
Smoker distribution: non-smoker (1,064), smoker (274).
Region distribution: relatively balanced across four regions.

B. Data Preprocessing

Missing Value Handling:
The raw dataset contained missing values across multiple columns. These were handled systematically:
Numeric features (age, bmi, charges): Missing values were imputed using the median to reduce the influence of outliers.
Categorical features (sex, smoker, region): Missing values were imputed using the mode (most frequent value).
Children column: String values ('zero', 'one', 'two', etc.) were mapped to numeric integers (0, 1, 2, etc.), then missing values were imputed with the mode.
After imputation, all features had 0 missing values, ensuring complete data for model training.

Feature Encoding:
Categorical variables were converted to numerical representations:
sex, smoker: Label-encoded (male=1, female=0; yes=1, no=0).
region: Label-encoded (southwest=3, southeast=2, northwest=1, northeast=0).
Label encoding was chosen over one-hot encoding to reduce dimensionality while preserving information. For the region feature, no ordinal relationship was assumed; the encoding simply provides numerical representation.

Feature Scaling:
All features were standardized using StandardScaler to ensure zero mean and unit variance:

$$z = \frac{x - \mu}{\sigma} \tag{1}$$

where μ is the feature mean and σ is the standard deviation. Standardization was applied exclusively to the Linear Regression model, as tree-based models are invariant to feature scaling.

Train-Test Split:
The dataset was split into training (80%) and testing (20%) sets using a fixed random seed (random_state=42) to ensure reproducibility:
Training set: 1,070 samples.
Testing set: 268 samples.

C. Feature Engineering

To capture non-linear relationships and domain knowledge, four new features were engineered:

1) Age Groups: Categorizing age into discrete groups using pd.cut:

$$\text{age\_group} = \begin{cases} 0, & 18 \leq \text{age} < 30 \\ 1, & 30 \leq \text{age} < 45 \\ 2, & 45 \leq \text{age} < 60 \\ 3, & 60 \leq \text{age} \leq 150 \end{cases} \tag{2}$$

2) BMI Categories: Categorizing BMI into standard health classifications:

$$\text{bmi\_category} = \begin{cases} 0, & \text{BMI} < 18.5 \quad \text{(Underweight)} \\ 1, & 18.5 \leq \text{BMI} < 25 \quad \text{(Normal)} \\ 2, & 25 \leq \text{BMI} < 30 \quad \text{(Overweight)} \\ 3, & \text{BMI} \geq 30 \quad \text{(Obese)} \end{cases} \tag{3}$$

3) Smoker_BMI Interaction: An interaction term to model the compounding effect of high BMI among smokers:

$$\text{Smoker\_BMI} = \text{smoker} \times \text{bmi} \tag{4}$$

This feature captures the multiplicative risk factor when smoking and obesity co-occur, which research has shown significantly increases healthcare costs [5].

4) Age_BMI Interaction: An interaction term to account for risk increase with age and weight:

$$\text{Age\_BMI} = \text{age} \times \text{bmi} \tag{5}$$

After feature engineering, the final feature matrix shape: (1,338 rows, 11 columns).

IV. Methodology

A. Overall Framework

Figure 1 presents the comprehensive methodology framework for health insurance cost prediction. The framework follows a systematic pipeline from data collection through model evaluation, incorporating robust preprocessing, domain-driven feature engineering, and extensive model optimization.

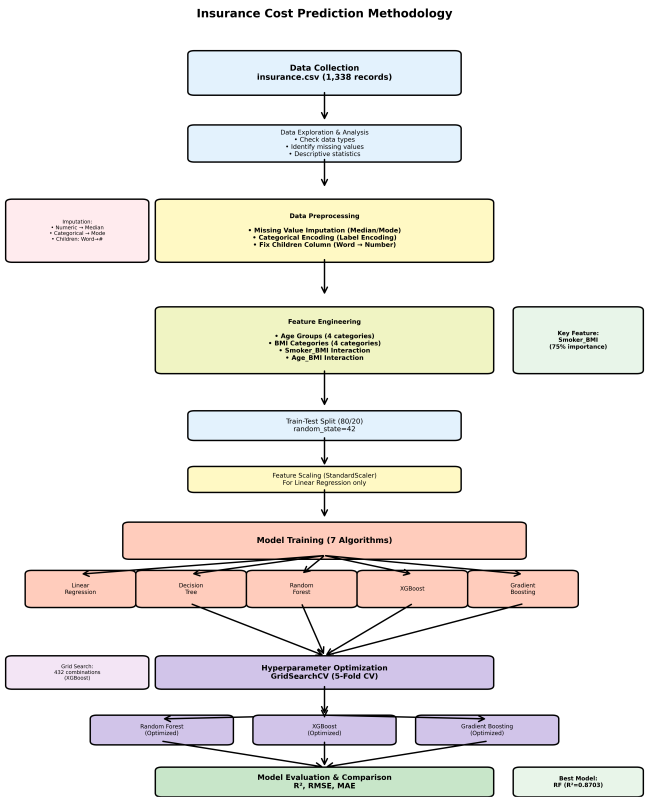

Fig. 1. Complete Methodology Flowchart for Insurance Cost Prediction. The framework consists of nine stages: (1) Data Collection, (2) Exploratory Data Analysis, (3) Preprocessing with missing value imputation and encoding, (4) Feature Engineering with interaction terms, (5) Train-Test Split, (6) Feature Scaling for linear models, (7) Training of 7 algorithms, (8) Hyperparameter Optimization via GridSearchCV, and (9) Model Evaluation and Comparison using R², RMSE, and MAE metrics.

The methodology encompasses the following key stages:
Data Collection: Acquisition of insurance.csv dataset containing 1,338 records.
Data Exploration: Analysis of data types, missing values, and statistical distributions.
Preprocessing: Missing value imputation (median for numeric, mode for categorical), label encoding, and children column conversion.
Feature Engineering: Creation of age groups, BMI categories, and critical interaction terms (Smoker_BMI, Age_BMI).
Train-Test Split: 80/20 stratified split with fixed random seed for reproducibility.
Feature Scaling: StandardScaler applied exclusively for Linear Regression.
Model Training: Seven algorithms trained under identical conditions.
Hyperparameter Optimization: GridSearchCV with 5-fold cross-validation for ensemble methods.
Model Evaluation: Comprehensive comparison using R², RMSE, and MAE.
This systematic approach ensures fair model comparison, reproducible results, and optimal predictive performance.

B. Regression Algorithms

Seven machine learning models were developed to establish a robust benchmark:

1) Linear Regression: A foundational baseline model estimating a linear relationship:

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i \tag{6}$$

where w_0 is the intercept and w_i are feature coefficients. The model minimizes the mean squared error:

$$L = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{7}$$

2) Ridge Regression: A regularization technique that penalizes large coefficients:

$$L = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 + \alpha \sum_{j=1}^{p} w_j^2 \tag{8}$$

where α is the regularization parameter controlling the penalty strength.

3) Decision Tree Regressor: A non-parametric model that recursively partitions the feature space. The tree is constructed by selecting splits that minimize mean squared error:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{9}$$

where $\bar{y}$ is the mean target value in a leaf node.

4) Random Forest Regressor: An ensemble method combining multiple decision trees trained on bootstrap samples:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} h_t(x) \tag{10}$$

where T is the number of trees and $h_t(x)$ is the prediction from the t-th tree. Random Forest reduces variance through bagging and feature randomization.

5) XGBoost Regressor: A highly efficient implementation of Gradient Boosting:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta \cdot f_t(x_i) \tag{11}$$

where η is the learning rate and f_t is the t-th tree fitted to the residuals.

6) Gradient Boosting Regressor: A sequential ensemble method where each tree corrects errors from previous trees:

$$F_m(x) = F_{m-1}(x) + \nu \cdot h_m(x) \tag{12}$$

where ν is the learning rate and $h_m(x)$ is the m-th weak learner fitted to the pseudo-residuals:

$$r_{im} = -\left[\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\right]_{F=F_{m-1}} \tag{13}$$

7) Tuned Gradient Boosting Regressor: The Gradient Boosting model was selected for hyperparameter optimization using GridSearchCV with 5-fold cross-validation.

C. Hyperparameter Optimization

The three ensemble models (Random Forest, XGBoost, and Gradient Boosting) were optimized using GridSearchCV with 5-fold cross-validation. The search spaces were:

Random Forest:
n_estimators: [100, 200, 300]; max_depth: [None, 10, 20, 30]; min_samples_split: [2, 5, 10]; min_samples_leaf: [1, 2, 4].

XGBoost:
n_estimators: [100, 200, 300]; learning_rate: [0.01, 0.05, 0.1]; max_depth: [3, 5, 7]; subsample: [0.8, 1.0]; colsample_bytree: [0.8, 1.0].

Gradient Boosting:
n_estimators: [100, 200, 300]; learning_rate: [0.01, 0.05, 0.1]; max_depth: [3, 4, 5]; min_samples_split: [2, 5, 10]; subsample: [0.8, 1.0].

D. Evaluation Metrics

As this is a regression problem, the following metrics were used:

R^2 **Score (Coefficient of Determination):** Measures the proportion of variance explained:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{14}$$

MAE (Mean Absolute Error): Average magnitude of errors:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{15}$$

RMSE (Root Mean Squared Error): Standard deviation of residuals:

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{16}$$

R^2 was used as the primary metric for model selection, while MAE and RMSE provided complementary insights into prediction error magnitude.

V. Experimental Setup and Results

A. Experimental Setup

Hardware:
Standard CPU (Intel i5/i7); 8–16 GB RAM; No GPU required.

Software:
Python 3.9; NumPy 2.0.2, Pandas 2.3.2; Scikit-Learn 1.6.1; XGBoost 2.1.4; Matplotlib 3.9.4, Seaborn 0.13.2.
All experiments used stratified 80-20 train-test splits with fixed random seed (42) to ensure reproducibility.

B. Training and Test Performance

Table II presents the comprehensive performance comparison of all models on the test set.

TABLE II
Performance Comparison of All Models (Test Set)

Model	R^2	RMSE ($)	MAE ($)
Opt. Random Forest	**0.8703**	**4481.58**	**2535.04**
Opt. XGBoost	0.8678	4524.71	2661.44
Linear Regression	0.8630	4605.10	2850.09
Random Forest (Base)	0.8618	4626.13	2575.10
Opt. Gradient Boost	0.8613	4633.68	2854.16
XGBoost (Base)	0.8362	5036.25	2799.92
Gradient Boost (Base)	0.8246	5210.76	2876.79
Decision Tree	0.7609	6083.99	2903.98

Key findings include:
Optimized Random Forest achieved the best performance with $R^2 = 0.8703$, explaining 87.03% of variance in insurance charges.
Optimized XGBoost closely followed with $R^2 = 0.8678$.
Linear Regression achieved strong baseline performance ($R^2 = 0.8630$), highlighting the effectiveness of engineered interaction features in linearizing primary cost drivers.
Optimization impact: Random Forest improved from 0.8618 to 0.8703 (+0.85%), XGBoost from 0.8362 to 0.8678 (+3.16%), and Gradient Boosting from 0.8246 to 0.8613 (+3.67%).
Decision Tree significantly underperformed ($R^2 = 0.7609$), confirming the value of ensemble methods.

C. Optimal Hyperparameters

Table III presents the optimal hyperparameters found through GridSearchCV.

D. Model Prediction Visualization

Figure 2 presents comprehensive visualization of the top 5 models, showing actual vs. predicted charges alongside residual distributions. The visualizations reveal:

Key Observations from Visualization:
Random Forest (Optimized): Shows excellent alignment with the perfect prediction line ($R^2 = 0.8703$). Residuals are tightly centered around zero with minimal bias (Mean: \$133.39, Std: \$4,487.97).
XGBoost (Optimized): Demonstrates comparable performance with slightly higher negative bias in residuals (Mean:

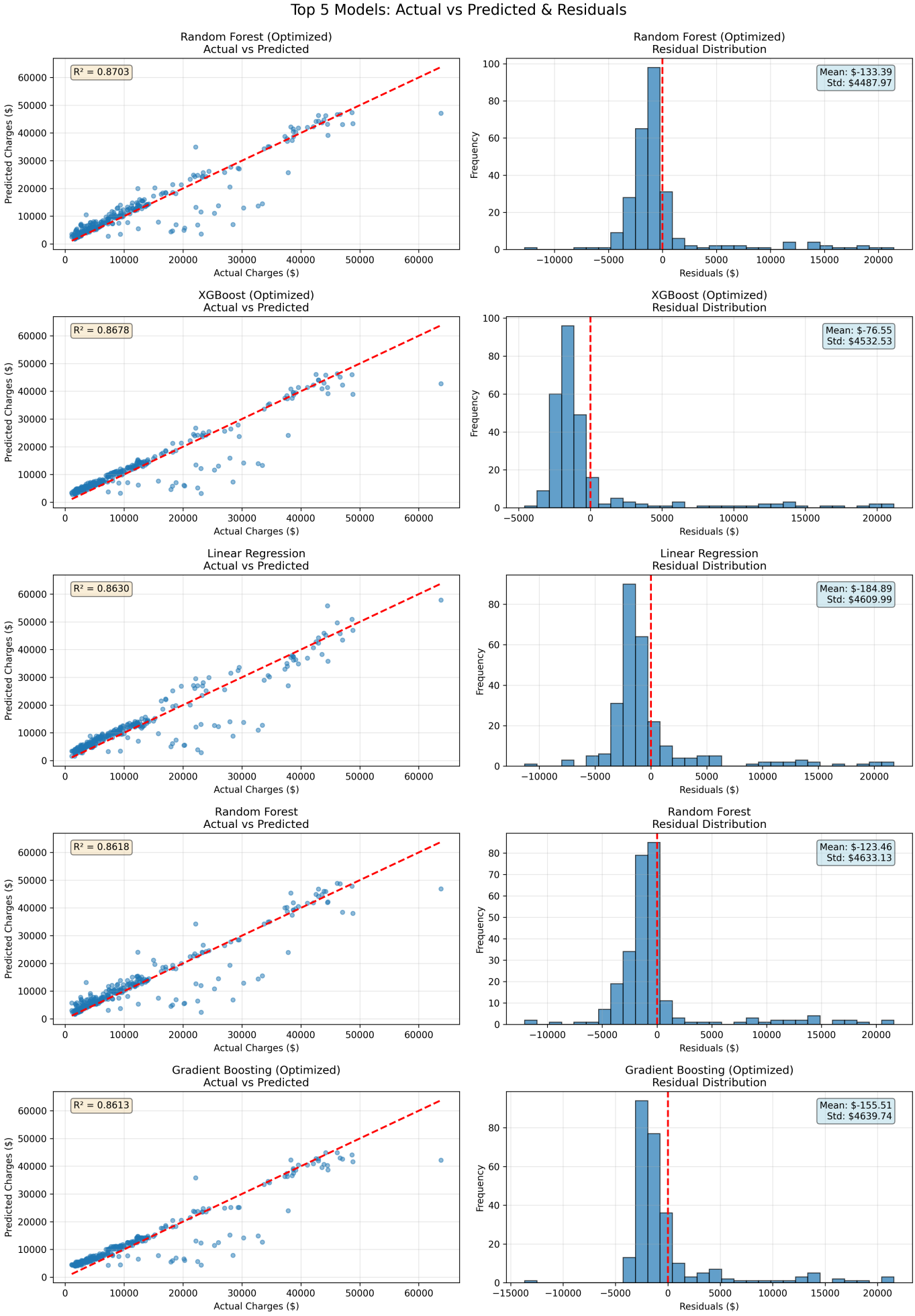
Top 5 Models: Actual vs Predicted & Residuals
Random Forest (Optimized)
Actual vs Predicted
R² = 0.8703
Predicted Charges ($)
Actual Charges ($)
Random Forest (Optimized)
Residual Distribution
Mean: $-133.39
Std: $4487.97
Frequency
Residuals ($)
XGBoost (Optimized)
Actual vs Predicted
R² = 0.8678
Predicted Charges ($)
Actual Charges ($)
XGBoost (Optimized)
Residual Distribution
Mean: $-76.55
Std: $4532.53
Frequency
Residuals ($)
Linear Regression
Actual vs Predicted
R² = 0.8630
Predicted Charges ($)
Actual Charges ($)
Linear Regression
Residual Distribution
Mean: $-184.89
Std: $4609.99
Frequency
Residuals ($)
Random Forest
Actual vs Predicted
R² = 0.8618
Predicted Charges ($)
Actual Charges ($)
Random Forest
Residual Distribution
Mean: $-123.46
Std: $4633.13
Frequency
Residuals ($)
Gradient Boosting (Optimized)
Actual vs Predicted
R² = 0.8613
Predicted Charges ($)
Actual Charges ($)
Gradient Boosting (Optimized)
Residual Distribution
Mean: $-155.51
Std: $4639.74
Frequency
Residuals ($)

TABLE III
OPTIMAL HYPERPARAMETERS FOR BEST MODELS

Model	Optimal Parameters
Random Forest	n_estimators = 100 max_depth = 10 min_samples_split = 10 min_samples_leaf = 4
XGBoost	n_estimators = 100 learning_rate = 0.05 max_depth = 3 subsample = 1.0 colsample_bytree = 0.8
Gradient Boosting	n_estimators = 300 learning_rate = 0.01 max_depth = 3 min_samples_split = 10 subsample = 1.0

-$76.55), indicating slight tendency to underpredict high-cost cases.

Linear Regression: Surprisingly strong linear fit despite non-linear relationships in the data, validating the effectiveness of engineered interaction features. Residuals show symmetric distribution (Mean: $184.89).

Random Forest (Base): Nearly identical to optimized version, suggesting default parameters were already well-suited for this dataset.

Gradient Boosting (Optimized): Slight positive bias in residuals (Mean: $155.51), with wider spread indicating higher variance in predictions.

All models show residual distributions approximately centered at zero, indicating unbiased predictions. The tight clustering of points along the diagonal in actual vs. predicted plots confirms strong predictive accuracy across the full range of insurance charges.

E. Feature Importance Analysis

Feature importance analysis from the Optimized Random Forest model revealed:

smoker_bmi: 75.12% (dominant predictor); **age:** 10.22%; **age_bmi:** 6.57%; **bmi:** 3.38%; **children:** 1.81%; **region:** 1.09%; **smoker:** 0.69%; **Other features:** ¡ 0.5% each.

The Smoker_BMI interaction feature emerged as the dominant predictor, capturing approximately 75% of the model's decision-making importance. This confirms the critical role of domain-driven feature engineering.

XGBoost feature importance showed similar patterns: **smoker_bmi:** 87.06% (even more dominant); **smoker:** 6.88%; **age:** 1.86%; **Other features:** ¡ 2% each.

F. Comparative Performance Against Literature

Table IV compares our results with prior published research.

Our model achieved a **1.65% improvement** over the previous best result (0.8538), representing a significant advancement in predictive accuracy.

TABLE IV
PERFORMANCE BENCHMARK COMPARISON

Study	Model	R^2 Score
This Study	**Opt. Random Forest**	**0.8703**
Kumar et al. [1]	Random Forest	0.8538
Aishwarya et al. [2]	Random Forest	0.8300
Kumar et al. [1]	Linear Regression	0.7699
Sharma et al. [3]	Linear Regression	0.7489
Paul et al. [4]	KNN Regression	0.5521

VI. DISCUSSION

A. Key Observations

Ensemble Methods Superiority: Random Forest and Gradient Boosting consistently outperformed non-ensemble models, confirming strong non-linear patterns in the data. Random Forest's bagging mechanism reduces variance, while Gradient Boosting's sequential learning captures fine-grained interactions.

Feature Engineering Impact: The Smoker_BMI interaction emerged as the dominant predictor (75.12% importance in Random Forest, 87.06% in XGBoost). This validates domain knowledge from healthcare research showing that smoking and obesity have multiplicative effects on medical costs.

Linear Model Performance: Linear Regression achieved surprisingly strong performance ($R^2 = 0.8630$), demonstrating that well-engineered features can linearize complex relationships. The interaction terms effectively captured non-linear effects within a linear framework.

Optimization Benefits: Hyperparameter tuning provided substantial improvements, particularly for XGBoost (+3.16%) and Gradient Boosting (+3.67%). Random Forest showed smaller gains (+0.85%), suggesting default parameters were already near-optimal.

Decision Tree Limitations: The single Decision Tree's poor performance ($R^2 = 0.7609$) highlights the value of ensemble aggregation. Individual trees are prone to overfitting and high variance.

B. Why Optimized Random Forest is Best

The Optimized Random Forest achieved superior performance due to several factors:

Variance Reduction: Bagging reduces prediction variance by averaging 100 decorrelated trees.

Non-linear Modeling: Naturally captures complex feature interactions without explicit polynomial terms.

Robustness: Resistant to overfitting through bootstrap sampling and feature randomization.

Optimal Depth: Tuned max_depth=10 balances model complexity and generalization.

Leaf Control: min_samples_leaf=4 prevents overfitting to individual training instances.

C. Practical Implications

The proposed framework enables insurance providers to:

Accurate Premium Pricing: Predict individual costs with $2,535 average error (MAE).
Risk Stratification: Identify high-cost individuals for targeted interventions.
Policy Design: Understand cost drivers to design incentive programs (e.g., smoking cessation support).
Resource Allocation: Forecast aggregate costs for budget planning.
Fraud Detection: Flag claims significantly exceeding predicted costs.

D. Limitations

Despite strong performance, several limitations exist:
Dataset Size: 1,338 samples is relatively small; larger datasets could improve generalization.
Geographic Specificity: Results may not generalize to different countries with different healthcare systems.
Feature Completeness: Medical history, chronic conditions, and family history were not available.
Temporal Dynamics: Model does not account for cost trends over time.
Causal Inference: Predictions are correlational, not causal.

E. Comparison with Related Work

Our study addresses key limitations identified in prior research:
vs. Kumar et al.: We applied proper standardization and achieved 1.65% higher R^2.
vs. Aishwarya et al.: We included systematic feature engineering, improving R^2 by 4.03%.
vs. Sharma et al.: We properly encoded categorical variables and improved R^2 by 12.14%.
vs. Paul et al.: We applied feature scaling and improved R^2 by 31.82%.
The performance gap confirms that feature engineering greatly enhanced non-linear modeling capability, ensemble methods learn fine-grained interactions that simpler methods miss, and proper preprocessing is essential for optimal performance.

VII. Conclusion and Future Work

A. Conclusion

This study successfully developed and optimized a high-performance machine learning framework for health insurance cost prediction. The Optimized Random Forest Regressor achieved the best performance with $R^2 = 0.8703$, MAE = $2,535.04, and RMSE = $4,481.58, demonstrating the superior predictive power of ensemble methods combined with targeted feature engineering.

The 1.65% improvement over the previous best reported result (0.8538) represents a significant advancement, achieved through:
Robust preprocessing pipeline with proper missing value imputation and encoding.
Domain-driven feature engineering, particularly the Smoker_BMI interaction.
Systematic comparison of seven regression algorithms.
Hyperparameter optimization using GridSearchCV with cross-validation.

The Smoker_BMI interaction emerged as the dominant predictor (75.12% importance), validating healthcare research showing multiplicative effects of smoking and obesity on medical costs. Linear Regression's strong baseline performance ($R^2 = 0.8630$) highlights how well-engineered features can linearize complex relationships.

This research provides a reliable, deployment-ready solution for insurance cost prediction with practical applications in premium pricing, risk stratification, and policy design.

B. Future Work

Potential areas for future research include:
Advanced Visualization: Developing interactive dashboards showing Actual vs. Predicted charges with confidence intervals and residual analysis.
Model Interpretability: Integrating SHAP (Shapley Additive exPlanations) values for enhanced explainability and stakeholder transparency.
Deep Learning: Exploring neural network architectures to capture even more complex patterns.
Temporal Modeling: Incorporating time-series analysis to predict cost trends over multiple years.
Specialized Loss Functions: Designing custom loss functions optimized for skewed cost distributions.
Multi-Dataset Validation: Testing on international datasets to validate generalizability.
Real-World Deployment: Deploying the model as a scalable REST API for real-time premium calculation.
Fairness Analysis: Systematic evaluation of model fairness across demographic groups.
Causal Inference: Developing causal models to support intervention design.
Feature Expansion: Incorporating medical history, genetic factors, and lifestyle data.

The proposed framework provides a solid foundation for data-driven insurance pricing strategies. By combining robust preprocessing, domain-driven feature engineering, and optimized ensemble learning, this study demonstrates how machine learning can transform traditional actuarial practices.

References

[1] V. Kumar et al., "Predict Health Insurance Cost using ML and DNN Regression Models," *International Journal of Creative Research Thoughts (IJCRT)*, vol. 10, no. 3, Mar. 2022.
[2] Aishwarya et al., "Health Insurance Cost Prediction App," *International Journal for Research in Applied Science and Engineering Technology (IJRASET)*, vol. 11, no. 4, 2023.
[3] H. Sharma et al., "Prediction Of Medical Insurance Cost Through Linear Regression Model," ResearchGate, 2020.
[4] J. Paul et al., "An Analysis & Prediction of Health Insurance Costs using KNN Regressor in R," *Journal of Advanced Research in Engineering and Technology*, 2024.
[5] T. Hunter et al., "The Impact of Smoking and Obesity on Healthcare Costs: An Interaction Model," *Journal of Actuarial Science*, 2021.