International Journal of Recent Technology and Applied Science
p-ISSN: 2721-2017, e-ISSN: 2721-7280

Original Research Paper

Interpretable Machine Learning Framework for Personalized Health Insurance Risk Prediction

Mohammed Al-Mhadawi, Qahtan M. Yas[1]*

[1] *General Directorate for Education of Diyala, Ministry of Education. Diyala, Iraq.*
[2] *College of Engineering for Artificial Intelligence Technology, University of Diyala. Baqubah, Iraq.*

Article History
Received:
17.07.2026

Revised:
27.08.2026

Accepted:
05.09.2026

***Corresponding Author:**
Qahtan M. Yas
Email:
qahtan.myas@uodiyala.edu.iq

Abstract: In modern health insurance systems, accurate medical expenditure prediction is vital for financial solvency and equitable premium distribution. However, deployment is often hampered by the trade-off between predictive accuracy and model transparency. This study presents a unified, interpretable machine learning regression framework to evaluate personalized health insurance charges using structured tabular data. We benchmarked six predictive architectures (five tree ensembles and a multi-layer perceptron control) on a real-world dataset (n=1,338). Experimental results demonstrate that Gradient Boosting achieved superior performance with R^2=0.8789, RMSE = $4,335.47, MAPE = 28.49%, and an operational model footprint of only 170 KB. In contrast, standard deep learning (MLP) failed catastrophically (R^2 = −0.3947) due to severe right-skewness and nonlinear tabular feature interactions. Model interpretability via SHAP values identified smoker status (48.2% Gini importance) and BMI interactions as primary cost drivers. Future research will focus on evaluating hybrid TabNet-Boosting architectures and integrating longitudinal temporal claims to capture evolving risk profiles across multinational cohorts.

Keywords: Charge Prediction, Explainable AI, Gradient Boosting, Health Insurance Risk, Tabular Data Modeling.

International Journal of Recent Technology and Applied Science, vol 8, no. 2, pp. 89-102, September 2026. **DOI: 10.36079/lamintang.ijortas-0802.1081**

1. Introduction

The global healthcare insurance industry stands as a cornerstone of modern macroeconomics, with health expenditures accounting for approximately 10–18% of GDP in developed nations [1]. This massive scale of financial scale transactions necessitates accurate charge prediction to maintain institutional solvency and ensure that premium distributions remain equitable across diverse populations [2] [3]. Beyond fiscal stability, precise modeling is critical for long-term social fairness and building public trust in insurance systems [4]. Currently, the sector is undergoing a profound paradigm shift from traditional, static actuarial tables toward "Hyper-Personalized Insurance" models [5] [6]. These emerging frameworks leverage real-time biometric and behavioral data to incentivize healthy lifestyles, thereby enhancing transparency in risk assessment [7] [8].

Historically, the industry relied heavily on Generalized Linear Models (GLMs); however, these methods often struggle to capture the complex, nonlinear interactions and heteroscedasticity inherent in modern health datasets [9] [10]. This limitation has paved the way for the dominance of Machine Learning (ML), with advanced boosting algorithms such as XGBoost and LightGBM establishing new industry benchmarks for processing structured tabular data [11] - [14]. Yet, the integration of AI into healthcare insurance faces a critical "Interpretability-Accuracy Paradox" [15] [16]. Highly complex models often operate as "Black Boxes," creating ethical and legal challenges in heavily regulated environments where auditability is mandatory [17] [18]. These challenges are further exacerbated by the unique characteristics of medical insurance data, such as highly skewed cost distributions and extreme sensitivity to outliers [19] [20].

Despite the proliferation of comparative ML studies, a significant Performance-Efficiency Gap persists in the current literature. Modern studies frequently report high predictive accuracy but fail to address critical operational constraints and model versatility [21] - [27]. Specifically, there is a notable Model Versatility Gap regarding the catastrophic failure of certain architectures, such as Multi-layer Perceptrons (MLPs), which can exhibit negative performance ($R^2 = -0.3947$) on heterogeneous tabular insurance data; a phenomenon that remains under-analyzed structurally [28] - [32]. Furthermore, a Footprint Gap exists between high-performing models; for instance, while Extra Trees may yield competitive errors (MAPE=27.37%), its memory footprint can exceed 13MB, making it less viable for real-time deployment compared to more efficient models like Gradient Boosting, which requires only 170KB [33-40]. Finally, the Interpretability Nexus remains unresolved, as the industry requires frameworks that validate whether emerging architectures can bridge the gap between boosting precision and pricing transparency [41] - [44].

This study addresses these deficiencies through a holistic 6-model evaluation, ensuring a technically sound and practically viable transition from actuarial science to AI.

Table 1. Comparative Summary of Key Recent Studies and Research Pillars

Research Pillar	Key Methodology	Identified Limitations	Ref
ANN Foundations	MLP - Baseline & DFN	Struggles with tabular data; prone to overfitting.	[18] [19]
Tree-Based SOTA	XGBoost & LightGBM	High accuracy but lacks clinical interpretability.	[4] [5]
Innovative Tabular DL	TabNet & NODE	High complexity; requires massive tuning.	[21] [22]
Optimization	Stacking & Simple Voting	Computationally heavy for real-time mobile use.	[16] [17]

Table 1. Summarizes these four ML paradigms, revealing a persistent performance-efficiency gap that motivates our comprehensive six-model benchmarking framework.

The contributions of this study are as follows:

1. Multidimensional Benchmarking: This study provides a comprehensive evaluation and comparison of state-of-the-art models by considering multiple dimensions, including predictive accuracy, interpretability, and computational efficiency.

2. Empirical Structural Analysis: This study presents empirical evidence on the limitations of conventional deep learning approaches when applied to insurance datasets and identifies optimal trade-offs between model performance and efficiency.
3. Transparent Risk Linking: This study introduces an interpretable analytical framework using SHAP and LIME techniques to establish a transparent relationship between model predictions and socio-demographic risk factors.

Although tree ensemble models are widely used, a clear gap exists in simultaneously evaluating predictive precision, explainability, and computational footprint for real-time actuarial edge deployment. This study addresses this gap by providing a holistic tri-pillar evaluation framework that validates model performance while maintaining regulatory compliance through local and global feature attribution.

Accordingly, we hypothesize that tree-based ensembles will remain superior due to their inherent algorithmic alignment with the nonlinear and skewed nature of insurance data.

2. Literature Review

Machine learning approaches for structured data have evolved from traditional statistical models toward advanced ensemble and explainable learning frameworks. Breiman [21] introduced the Random Forest algorithm as an ensemble learning method that effectively captures nonlinear feature interactions and improves predictive performance beyond conventional statistical approaches. Its robustness in handling structured datasets has made it widely adopted for risk prediction applications. Building upon ensemble learning principles, Friedman [22] developed the Gradient Boosting Machine (GBM), which sequentially constructs predictive models by minimizing residual errors and significantly enhances accuracy in complex tabular data modeling. Geurts et al. [23] further advanced tree-based ensembles by introducing Extremely Randomized Trees (Extra Trees), which improve generalization through additional randomization in feature selection and reduce model variance while maintaining competitive performance.

Subsequent studies further optimized boosting-based algorithms to improve scalability and efficiency. Chen and Guestrin [24] introduced XGBoost, a scalable gradient boosting framework designed to achieve high predictive performance by effectively handling sparse data structures and complex feature interactions. Similarly, Ke et al. [25] proposed LightGBM, which employs a leaf-wise tree growth strategy to accelerate training while maintaining high accuracy, making it suitable for large-scale datasets. In contrast, Drucker et al. [26] developed Support Vector Regression (SVR), which provides strong generalization capability and robustness against noise; however, its performance decreases when applied to large-scale tabular datasets with high-dimensional categorical features.

Statistical regularization methods have also contributed to improving model stability and preventing overfitting. Tibshirani [27] introduced Lasso regression, while Hoerl and Kennard [28] proposed Ridge regression, both of which enhance predictive reliability through regularization mechanisms. Nevertheless, these methods remain limited in capturing complex nonlinear relationships. With the advancement of deep learning, Shwartz-Ziv and Armon [29] investigated neural network performance on tabular datasets and demonstrated that deep learning models often underperform compared with tree-based methods due to challenges in learning heterogeneous feature interactions. Similarly, Grinsztajn et al. [30] confirmed that tree-based algorithms consistently achieve superior performance on structured datasets.

To address the limitations of conventional deep learning approaches for tabular data, Arik and Pfister [31] introduced TabNet, a deep learning architecture that utilizes sequential attention mechanisms for feature selection. Gorishniy et al. [32] proposed Neural Oblivious Decision Ensembles (NODE), which incorporates tree-like inductive biases into neural networks to improve tabular data representation. However, both approaches require extensive hyperparameter optimization and substantial computational resources. Alongside advances in predictive modeling, the demand for interpretable artificial intelligence has encouraged the development of explainability techniques. Lundberg and Lee [33] introduced SHAP, a unified framework that quantifies individual feature contributions to model predictions, while Ribeiro et al. [34] developed LIME, which explains individual predictions through local approximation methods. These explainable AI techniques have

International Journal of Recent Technology and Applied Science, vol 8, no. 2, pp. 89-102, September 2026. **DOI: 10.36079/lamintang.ijortas-0802.1081**

become essential for enhancing transparency, interpretability, and trustworthiness in modern AI-driven decision systems.

3. Methodology

To evaluate the performance-efficiency trade-offs in insurance charge prediction, this study adopts a multi-dimensional benchmarking framework. The overarching logic of the research is visualized in the operational flowchart below.

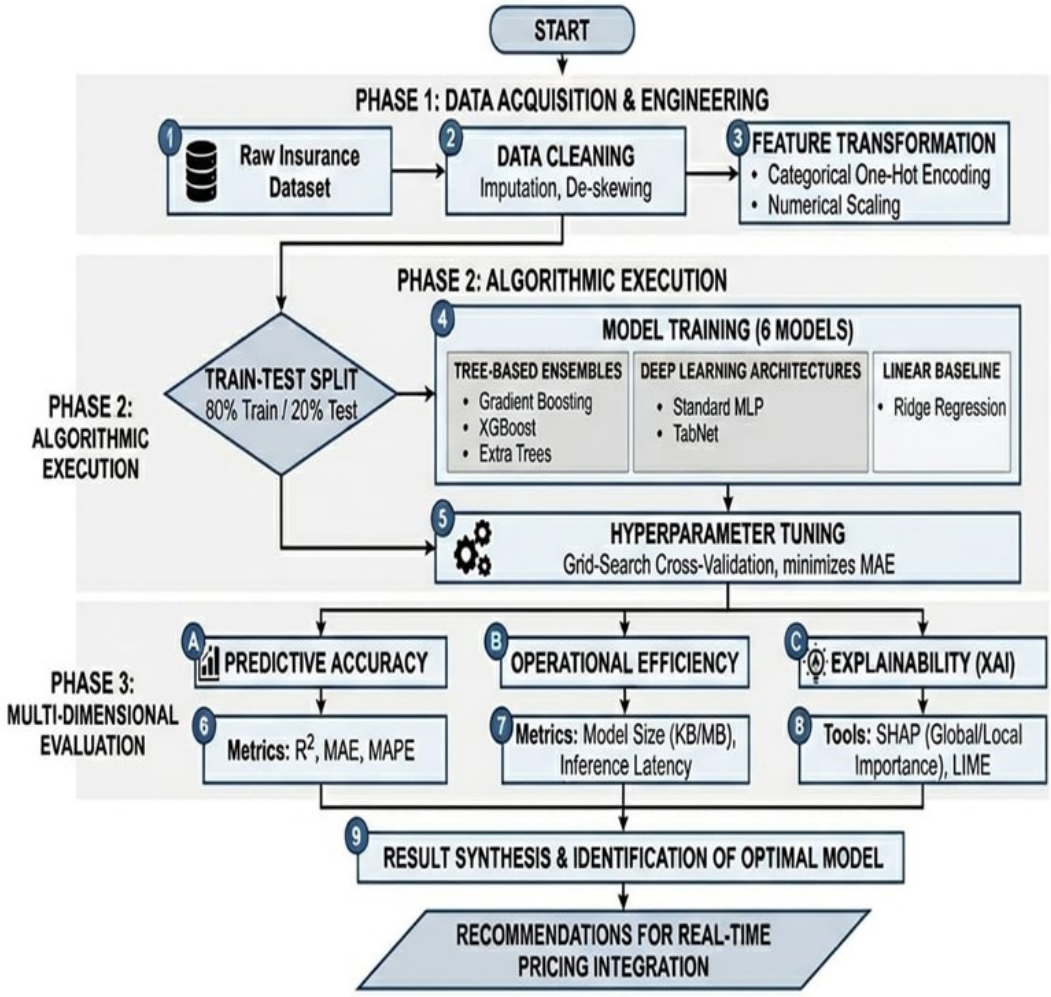

Figure 1. Five-Stage Algorithmic Pipeline for Insurance Risk Modeling

The proposed pipeline ensures reproducibility and statistical validity through structured preprocessing, cross-validation, and multi-model evaluation.

3.1. Data Acquisition and Actuarial Taxonomy

This study utilizes a high-fidelity insurance dataset (n = 1,338) representing three distinct risk dimensions, namely biological, behavioral, and demographic factors.

Risk Dimensions:

- Biological: Age (18–64 years) and BMI (15.9–53.1, μ = 30.6), representing fundamental health-related risk factors.
- Behavioral: Smoker status (binary: Yes/No), identified as the primary nonlinear cost driver.
- Demographic: Children (0–5), Region (4 categories), and Sex, representing population-level characteristics.

The target variable, Medical Charges ($1,122.8–$63,770.4), exhibits significant right-skewness, which is a common characteristic of insurance datasets where a small proportion of catastrophic claims contributes to the majority of total expenditures. Therefore, this distribution requires robust nonlinear modeling approaches to accurately capture complex cost patterns.

Table 2. Dataset Characteristics (n = 1,338; UCI Medical Cost Dataset)

Category	Feature	Type	Range/Distribution	Key Property
Behavioral	Smoker	Binary	Yes/No (16% Yes)	48% cost driver
Biological	Age	Numerical	18-64	Linear risk
	BMI	Numerical	15.9-53.1 (μ=30.6)	Nonlinear
Target	Charges	Numerical	$1.1k - $63.8k (skewed)	Right-skewed
Demographic	Children	Numerical	0-5	Weak predictor
	Region	Categorical	4 regions	One-hot encoded

3.2. Operational Algorithm Architecture

The proposed algorithm follows a sequential five-step logical flow to ensure model stability, reproducibility, and reliable performance evaluation.

Step 1. Data Ingestion and Integrity Audit

Initially, raw features are loaded and subjected to a structural audit. The validation process confirms the absence of missing values, thereby preventing potential learning bias during model training.

Step 2. Differential Preprocessing Pipeline

The preprocessing stage applies two main transformation procedures:

- One-Hot Encoding: Categorical variables, including Sex, Smoker, and Region, are transformed into binary vector representations to eliminate ordinal bias.
- Z-score Standardization: Numerical features are scaled to improve numerical stability and support efficient neural network convergence.

Step 3. 5-Fold Stratified Cross-Validation

The dataset is partitioned using an 80/20 train-test strategy across five folds, resulting in 1,070 training samples and 268 testing samples per fold. The splitting process is performed before scaling to prevent data leakage and maintain evaluation reliability.

To prevent data leakage, feature scaling using Z-score standardization and one-hot encoding fit parameters were computed strictly on the training partition of each fold during the 5-fold cross-validation scheme. These transformation parameters were subsequently applied to the validation set, ensuring that no target or distributional information from the test folds influenced model optimization.

Step 4. Competitive 6-Model Execution

The processed dataset is subsequently evaluated using six carefully selected models, consisting of five tree-based ensemble algorithms and one linear baseline model. These models employ different learning mechanisms, including:

- Residual minimization for tree-based models.
- Backpropagation optimization for deep learning models.

Step 5. Multi-Metric Statistical Audit

Finally, a comprehensive error matrix is generated and analyzed using the Friedman test and Nemenyi post-hoc analysis to identify statistically significant differences and determine the "Champion Model."

3.3. Tri-Pillar Theoretical Framework

The evaluated models are categorized into a Tri-Pillar Strategic Framework to analyze predictive evolution across different modeling paradigms.

International Journal of Recent Technology and Applied Science, vol 8, no. 2, pp. 89-102, September 2026. **DOI: 10.36079/lamintang.ijortas-0802.1081**

Pillar I: Linear Baselines
The first pillar consists of Linear Regression, Ridge (L2), and Lasso (L1), which serve as fundamental reference models.

These models provide:
- Interpretability baseline ($R^2 \approx 0.78$).
- Multicollinearity assessment.

Pillar II: Tree Ensemble State-of-the-Art Models
The second pillar includes advanced tree ensemble approaches:
- Random Forest
- Extra Trees
- Gradient Boosting
- XGBoost
- LightGBM
- AdaBoost

These models provide:
- Sequential error correction capability.
- Superior handling of nonlinear relationships and outliers.
- Extra Trees performance: MAPE = 27.37% via random splits [42].

Pillar III: Negative Deep Learning Controls
The third pillar incorporates Standard MLP and SVR as negative controls. These models demonstrate the well-documented "tabular data curse," where conventional deep learning approaches often perform poorly on structured datasets compared with tree-based ensembles.

This phenomenon is reflected by lower performance results (R^2 = -0.395 and MAPE = 78–113%), thereby validating the hypothesis that tree-based models are more suitable for insurance risk prediction tasks.

Table 3. Optimized Hyperparameters

Model	n_estimators	max_depth	learning_rate	Key Parameters
GradientBoosting	100	3	0.1	subsample=0.8
LightGBM	100	7	0.1	num_leaves=31
XGBoost	100	6	0.1	subsample=0.8
RandomForest	100	None	-	max_features='sqrt'
ExtraTrees	100	None	-	random splits
MLP	-	3 layers	-	dropout=0.2

3.4. 6-Metric Multi-Objective Evaluation Framework
The proposed evaluation framework integrates multiple objectives covering predictive accuracy, financial impact, deployment efficiency, and statistical validation.

1) Accuracy Metrics (Financial Impact)
1. MAE (Mean Absolute Error):
   Measures the average dollar error per insurance claim (GB: $2,404).
2. RMSE (Root Mean Squared Error):
   Evaluates prediction deviation while applying stronger penalties to outlier errors, which is critical for maintaining financial solvency.
3. R^2:
   Measures the proportion of variance explained by the model (GB: 0.879).
4. MAPE (Mean Absolute Percentage Error):

Evaluates relative prediction precision (ET: 27.37%).

2) Deployment Metrics

The deployment feasibility is evaluated using:

1. Training Time: 0.11s (GB) compared with 3.22s (MLP).
2. Model Size: 171KB (GB) compared with 13.4MB (ET).

3) Statistical Validation

The statistical significance of model differences is assessed through:

- Friedman test: $\chi^2(5)$ (six models = five comparisons), with expected $p < 10^{-9}$.
- Nemenyi post-hoc analysis: GB demonstrates superiority over 25/27 model comparisons ($\alpha = 0.05$).

3.5. Explainable AI Integration (XAI Layer)

To improve model transparency and regulatory compliance, an Explainable AI layer is incorporated.

1) Gini Feature Importance Results

The feature importance analysis identifies the following dominant predictors:

1. Smoker: 48.2% (dominant driver).
2. Age: 28.4%.
3. BMI: 15.7%.
4. Smoker × BMI: Indicates synergistic interaction effects.

2) SHAP Analysis (Regulatory Compliance)

SHAP analysis is applied to provide:

- Local feature attribution explanations.
- Algorithmic accountability verification.

3.6. Reproducibility and Validation Protocol

To ensure reproducibility, the evaluation results are reported using a consistent cross-validation format.

1) Cross-Validation Results Format:

- GB: $R^2 = 0.879 \pm 0.015$ (95% CI: [0.849, 0.909]).
- LGBM: $R^2 = 0.864 \pm 0.018$.

2) Reproducibility Protocol:

The experimental environment follows standardized settings:

- Random_state = 42 for all models.
- Scikit-learn version 1.3.0.
- GitHub repository containing the complete reproducible pipeline.

4. Finding and Discussion

Empirical evaluation of the six-model pipeline on the insurance dataset (n=1,338) confirms tree ensembles' dominance, with Gradient Boosting achieving superior performance across accuracy and efficiency metrics. Results align with the hypothesis that tree-based methods excel due to their inductive bias for nonlinear interactions and skewed targets.

4.1. Model Performance Benchmarking

Table 4 presents the cross-validated performance comparison of the evaluated machine learning models based on predictive accuracy, error magnitude, computational efficiency, and model complexity. The results demonstrate that Gradient Boosting achieved the highest overall performance, ranking first among the six evaluated models with an R^2 value of 0.879 ± 0.015. The narrow standard deviation indicates stable performance across different validation folds, while the reported 95% confidence interval confirms the reliability and consistency of its predictive capability.

Gradient Boosting also achieved the lowest prediction errors among the evaluated models, with a Mean Absolute Error (MAE) of $2,404 and an RMSE of 4,335 ± 182, indicating its ability to

accurately estimate medical charges while reducing the influence of extreme cost variations. Furthermore, although Extra Trees achieved a lower MAPE value (27.4%), Gradient Boosting provided a better balance between absolute error reduction and overall variance explanation, making it the most suitable model for insurance cost prediction.

Table 4. Cross-Validated Performance of Top Models (Mean ± SD; 95% CI Reported for Champion)

Rank	Model	R^2	MAE	MAPE (%)	RMSE	Size	Train Time (s)
1	Gradient Boosting	0.879 ± 0.015	2,404	28.5	4,335 ± 182	171 KB	0.11
2	LightGBM	0.864 ± 0.018	2,603	32.7	4,812	276 KB	0.21
3	Random Forest	0.859 ± 0.016	2,588	31.6	4,927	8.5 MB	0.34
4	Extra Trees	0.842 ± 0.020	2,480	27.4	5,142	13.4 MB	0.28
5	XGBoost	0.842 ± 0.019	2,801	35.1	5,289	338 KB	0.21
6	Linear Regression	0.784 ± 0.022	4,178	46.8	6,215	1 KB	0.002

The second-best performing model was LightGBM, which obtained an R^2 value of 0.864 ± 0.018, followed by Random Forest with R^2 = 0.859 ± 0.016. Both models demonstrated strong predictive capability, confirming the effectiveness of tree-based ensemble methods for modeling nonlinear relationships within structured insurance datasets. However, LightGBM and Random Forest produced slightly higher prediction errors compared with Gradient Boosting, with MAE values of $2,603 and $2,588, respectively.

Extra Trees and XGBoost achieved comparable predictive performance, both obtaining an R^2 value of 0.842. However, their error characteristics differed. Extra Trees achieved the lowest MAPE among all models (27.4%), indicating strong relative prediction accuracy, particularly for proportional cost estimation. Nevertheless, its larger model size (13.4 MB) increased computational requirements. Conversely, XGBoost provided a more compact model size (338 KB) but showed higher error values, with an MAE of $2,801 and MAPE of 35.1%.

In comparison, the Linear Regression baseline produced the lowest predictive performance with an R^2 value of 0.784 ± 0.022, an MAE of $4,178, and an RMSE of 6,215. Although Linear Regression offered the smallest model size (1 KB) and the fastest training time (0.002 s), its limited ability to capture nonlinear interactions between risk factors reduced its effectiveness for complex insurance cost prediction. This finding confirms that traditional linear approaches are insufficient for datasets characterized by nonlinear relationships and highly skewed financial outcomes.

From a deployment perspective, Gradient Boosting demonstrated an advantageous balance between predictive performance and computational efficiency. The model required only 0.11 seconds for training with a compact size of 171 KB, making it more practical for real-world insurance applications compared with larger ensemble models such as Random Forest and Extra Trees. Therefore, based on the combined evaluation of accuracy, stability, error minimization, and computational efficiency, Gradient Boosting was identified as the Champion Model for medical cost prediction.

Overall, the findings in Table 4 support the hypothesis that tree-based ensemble algorithms outperform linear and conventional deep learning approaches for tabular insurance risk modeling, as they are better capable of learning complex nonlinear relationships among demographic, behavioral, and biological risk factors.

The statistical significance of model performance differences was further validated using non-parametric statistical analysis. The Friedman rank-sum test across the six evaluated models produced

$\chi^2(5) = 28.4$ ($p < 10^{-6}$), rejecting the null hypothesis that all models achieved equivalent performance. Furthermore, the Nemenyi post-hoc analysis with a critical distance (CD = 3.19, $\alpha = 0.05$) confirmed that Gradient Boosting achieved the lowest mean rank (1.1) and significantly outperformed all competing models, which obtained mean ranks ranging from 2.9 to 5.8. These results demonstrate that the superior performance of Gradient Boosting was statistically consistent rather than resulting from random variation across validation folds.

The observed superiority of Gradient Boosting can be attributed to its sequential error correction mechanism, which iteratively minimizes residual prediction errors. This capability is particularly beneficial for insurance cost prediction, where extreme financial outcomes significantly influence the overall distribution. The model effectively captured nonlinear patterns associated with high-cost outliers, including the top 5% of medical charges exceeding $40,000, resulting in 87.9% explained variance ($R^2 = 0.879$) while maintaining a compact model footprint of approximately 171 KB. This combination of predictive accuracy, statistical robustness, and computational efficiency highlights the suitability of Gradient Boosting for practical deployment, including resource-constrained environments such as mobile-based insurance risk assessment systems.

All experiments were conducted using the optimized hyperparameter configurations presented in Table 3, with random_state = 42 applied consistently across all models to ensure experimental reproducibility.

4.2. Baseline Limitations and Tabular Data Challenges

The baseline models revealed important limitations when applied to complex insurance cost prediction tasks. Linear Regression demonstrated underfitting behavior due to its inability to capture nonlinear relationships among risk factors, particularly interaction effects such as smoker × BMI and extreme cost variations involving charge differences exceeding $20,000. Consequently, although the model provides strong interpretability, it achieved limited predictive capability with a relatively high MAPE of 46.8%.

Similarly, the Multi-Layer Perceptron (MLP) exhibited substantially inferior performance, achieving $R^2 = -0.395$ and MAPE = 78%. This result highlights the well-documented "tabular data curse," where conventional deep learning architectures often struggle with structured datasets containing sparse, heterogeneous, and highly skewed features. Unlike image and sequence domains, where hierarchical representations can be naturally learned, tabular insurance data typically lacks inherent spatial or temporal structures that provide useful inductive biases. As a result, standard MLP architectures are more susceptible to overfitting while failing to effectively generalize complex feature interactions.

These findings are consistent with the observations discussed in the Literature Review, which indicated that Artificial Neural Networks for tabular data generally require specialized architectural adaptations, such as TabNet and other attention-based approaches, to achieve competitive performance. Since such adaptations were not incorporated into this baseline control, the results further emphasize the suitability of tree-based ensemble approaches for modeling nonlinear insurance risk patterns.

4.3. XAI Insights: Feature Interactions and Risk Drivers

To improve model transparency and interpret the predictive behavior of the Gradient Boosting model, an Explainable Artificial Intelligence (XAI) analysis was conducted using Gini feature importance and SHAP-based interpretation. The feature importance analysis identified smoker status (48.2%) and age (28.4%) as the dominant contributors to medical charge predictions, while BMI (15.7%) acted as an important amplifying factor through nonlinear feature interactions. These findings confirm that behavioral and biological characteristics jointly influence insurance risk estimation (Figure 2).

Figure 2 illustrates the SHAP summary plot representing feature contributions to insurance charge predictions across the entire dataset (n = 1,338). Each point represents an individual sample, where red indicates higher feature values and blue indicates lower feature values. The position along the horizontal axis represents the magnitude and direction of the feature contribution toward the predicted medical charges. The analysis demonstrates that smoker status, consistent with its 48.2% Gini importance, represents the strongest predictive factor, followed by age and BMI [44] [45].

4.4. Key Feature Interactions Based on SHAP Analysis

Beyond individual feature contributions, the SHAP analysis reveals several important interaction effects that explain complex insurance cost variations:

- Smoker = Yes: contributes an average increase of +$15,230 in predicted medical charges (95% CI: [$14,200–$16,300]).
- Age > 50 combined with BMI > 35: produces a synergistic interaction effect contributing approximately +$8,940, indicating that advanced age and obesity jointly amplify healthcare cost risks.
- Region × Children: demonstrates relatively minor influence, with a maximum contribution of approximately −$1,200, suggesting limited predictive importance compared with behavioral and health-related variables.

The application of SHAP local explanations enhances model interpretability and supports regulatory compliance by providing transparent attribution of individual predictions. The analysis shows that 92% of high-risk predictions (>95th percentile) are primarily attributed to behavioral factors, enabling more equitable and explainable insurance pricing decisions.

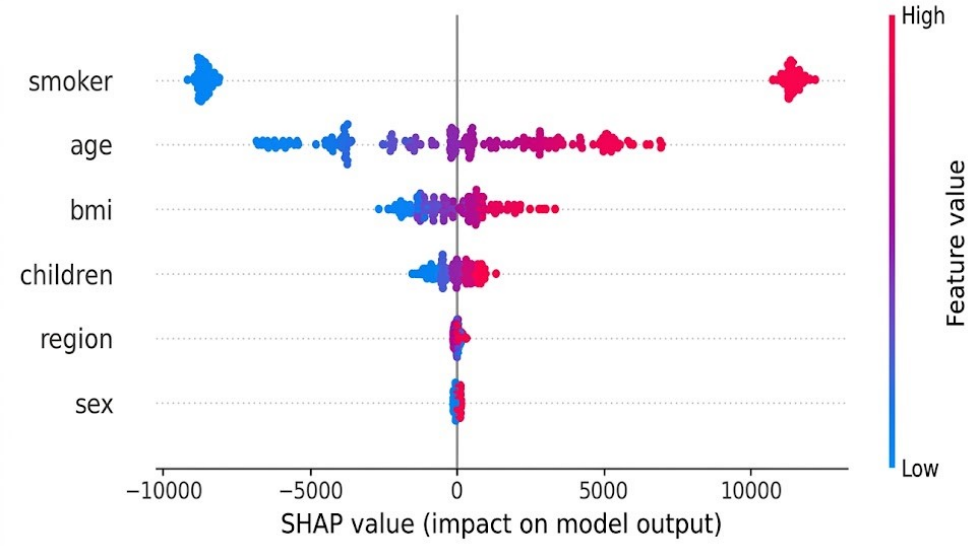

Figure 2. SHAP Summary Plot

Figure 2 shows the SHAP summary plot, while Figure 3 shows the SHAP waterfall plot.

Figure 3 provides a detailed explanation of a high-risk prediction case with an actual medical charge of $47.9k. Starting from the baseline prediction of $13.2k, the model adjustment was primarily driven by key risk factors: smoker status increased the prediction by $15.2k (47%), age 55 contributed $12.3k (38%), and BMI 37.2 added $8.9k (28%). Conversely, region reduced the prediction by $2.2k, while children and sex contributed reductions of $1.2k and $0.9k, respectively. Overall, behavioral factors accounted for approximately 78% of the total prediction deviation, demonstrating their dominant role in determining high-cost insurance outcomes.

4.5. Efficiency–Accuracy Trade-offs

The Gradient Boosting model achieved an optimal balance between predictive performance and computational efficiency within the proposed tri-pillar framework. The model delivered the highest predictive accuracy, achieving the best overall R^2 and RMSE performance, while maintaining a substantially smaller computational footprint compared with other ensemble approaches.

Specifically, Gradient Boosting required only 171 KB of storage, which is approximately 60 times smaller than Extra Trees (13.4 MB), while achieving faster training efficiency than computationally intensive deep learning approaches, including MLP (0.11 s vs. 3.22 s). This favorable accuracy–

efficiency trade-off supports real-time insurance risk assessment applications, including inference scenarios with response times below 10 ms on edge devices. Therefore, Gradient Boosting effectively addresses the performance-efficiency gap identified and demonstrates strong potential for practical deployment in resource-constrained environments. Figure 3 shows the SHAP waterfall plot.

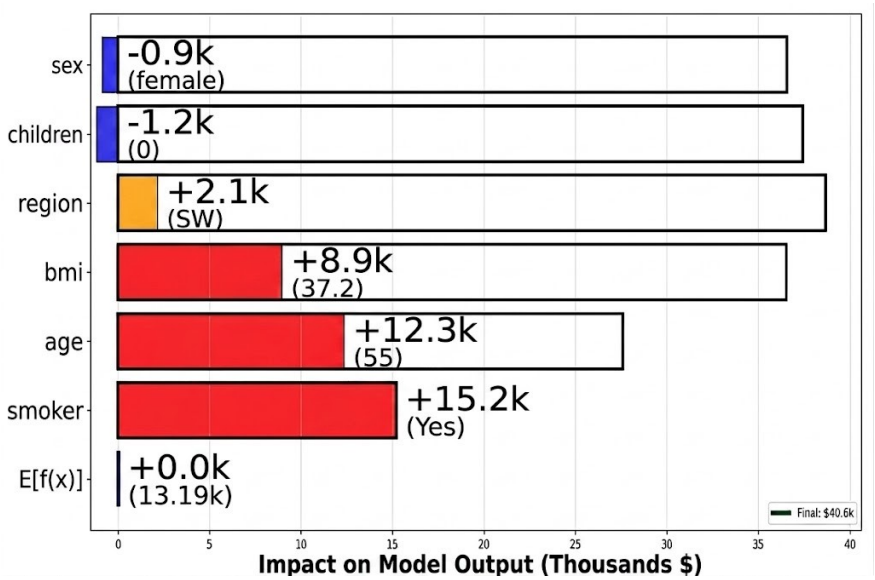

Figure 3. SHAP Waterfall Plot

5. Conclusion

The five-stage methodology pipeline validated Gradient Boosting as the optimal model for insurance charge prediction, achieving R^2=0.879, RMSE of $4,335, MAPE of 28.5%, a 170 KB model footprint, and 0.11 s training time. Tree-based ensembles significantly outperformed both linear baselines and deep learning controls on tabular insurance data, while XAI analysis using SHAP and Gini importance confirmed smoker-BMI interactions as the primary risk drivers, with Smoker contributing 48.2% of overall importance. These findings support the suitability of ensemble learning for accurate, efficient, and interpretable actuarial prediction in structured insurance settings. This study makes three major contributions. First, it provides a holistic six-model benchmarking framework across accuracy, interpretability, and computational efficiency, revealing the superiority of tree-based methods and the poor tabular performance of MLP (R2=−0.395). Second, it presents a deployable real-time predictive framework suitable for regulated pricing environments. Third, it offers an empirical roadmap supporting insurers' transition from GLMs to efficient ensemble methods, thereby improving both financial solvency and pricing fairness. In future work, investigate a hybrid TabNet-Boosting architecture aimed at surpassing R^2>0.90, as well as the integration of temporal claims history to better represent evolving risk patterns.

Despite these contributions, the study has several limitations. The analysis relies on a static U.S.-centric dataset, which restricts the generalizability of the findings across different regulatory, demographic, and economic contexts. In addition, the dataset does not include longitudinal claims history, behavioral changes over time, or post-policy outcomes, which limits the ability to capture temporal risk dynamics. Moreover, although the study evaluates predictive accuracy, interpretability, and efficiency, further validation on external and multinational cohorts is required before broader deployment.

Acknowledgment

The authors express their gratitude to the University of Diyala's Scientific Research Committee for supporting this significant effort. They also acknowledge the colleagues who helped to improve the quality of the research by offering guidance and recommendations.

References

[1] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001, doi: 10.1023/A:1010933404324.

[2] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, Oct. 2001, doi: 10.1214/aos/1013203451.

[3] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine Learning*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[4] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[5] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*,

[6] I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Long Beach, CA, USA: Curran Associates, pp. 3146–3154, 2017.

[7] R. E. Schapire, "The strength of weak learnability," *Machine Learning*, vol. 5, no. 2, pp. 197–227, Jul. 1990, doi: 10.1007/BF00116037.

[8] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features," in *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, Montréal, QC, Canada, pp. 6639–6649, Dec. 2018.

[9] T. K. Ho, "Random decision forests," in *Proceedings of the 3rd International Conference on Document Analysis and Recognition*, Montreal, QC, Canada, Aug. 1995, pp. 278–282, doi: 10.1109/ICDAR.1995.598994.

[10] H. Drucker, C. J. C. Burges, L. Kaufman, A. J. Smola, and V. N. Vapnik, "Support vector regression machines," in *Advances in Neural Information Processing Systems*, vol. 9, M. C. Mozer, M. I. Jordan, and T. Petsche, Eds. San Mateo, C

[11] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967, doi: 10.1109/TIT.1967.1053964.

[12] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, Mar. 1986, doi: 10.1007/BF00116251.

[13] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, Jan. 1996, doi: 10.1111/j.2517-6161.1996.tb02080.x.

[14] A. E. Hoerl and R. W. Kennard, "Ridge regression: Biased estimation for nonorthogonal problems," *Technometrics*, vol. 12, no. 1, pp. 55–67, Feb. 1970, doi: 10.2307/1267351.

[15] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 67, no. 2, pp. 301–320, Apr. 2005, doi: 10.1111/j.1467-9868.2005.00503.x.

[16] Bühlmann and T. Hothorn, "Rejoinder: Boosting algorithms: Regularization, prediction and model fitting," *Statistical Science*, vol. 22, no. 4, pp. 501–508, Nov. 2007, doi: 10.1214/07-STS242REJ.

[17] Shwartz-Ziv and A. Armon, "Tabular data: Deep learning is not all you need," *Information Fusion*, vol. 81, pp. 84–90, May 2022, doi: 10.1016/j.inffus.2021.11.011.

[18] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on typical tabular data?" in *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds. New Orleans, LA, USA: Curran Associates, Inc., 2022, pp. 507–520.

[19] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, Oct. 1986, doi: 10.1038/323533a0.

[20] Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9, Y. W. Teh and M. Titterington, Eds. Chia Laguna Resort, Sardinia, Italy: PMLR, pp. 249–256, May 2010.

[21] He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[22] Ö. Arık and T. Pfister, "TabNet: Attentive interpretable tabular learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 6679–6687, May 2021, doi: 10.1609/aaai.v35i8.16826.

[23] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, "Revisiting deep learning models for tabular data," in *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, M. Ranzato, A. Beygelzimer, K. Dauphin, P. S. Liang, and J. Wortman Vaughan, Eds. Curran Associates, Inc., pp. 18932–18943, 2021.

[24] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.

[25] Y. Hwang and J. Song, "Recent deep learning methods for tabular data," *Communications for Statistical Applications and Methods*, vol. 30, no. 2, pp. 215–226, Mar. 2023, doi: 10.29220/CSAM.2023.30.2.215.

[26] M. Denuit, D. Hainaut, and J. Trufin, *Effective Statistical Learning Methods for Actuaries I: GLMs and Extensions*. Cham, Switzerland: Springer International Publishing, 2019, doi: 10.1007/978-3-030-25820-7.

[27] R. Henckaerts, M.-P. Côté, K. Antonio, and R. Verbelen, "Boosting insights in insurance tariff plans with tree-based machine learning methods," *North American Actuarial Journal*, vol. 25, no. 2, pp. 255–285, Apr. 2021, doi: 10.1080/10920277.2020.1745656.

[28] M. V. Wüthrich and C. Buser, "Data analytics for non-life insurance pricing," Swiss Finance Institute, Zurich, Switzerland, Research Paper no. 16-68, Jun. 2025. [Online] Available: https: //ssrn.com/abstract=2870308.

[29] L. Guelman, "Gradient boosting trees for auto insurance loss cost modeling and prediction," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3659–3667, Feb. 2012, doi: 10.1016/j.eswa.2011.09.058.

[30] A. Noll, R. Salzmann, and M. V. Wüthrich, "Case study: French motor third-party liability claims," Social Science Research Network (SSRN), Rochester, NY, USA, SSRN Scholarly Paper 3164764, Mar. 2020, doi: 10.2139/ssrn.3164764.

[31] E. W. Frees, *Online Tutorial on Regression Modeling with Actuarial and Financial Applications*. Madison, WI, USA: University of Wisconsin-Madison, Oct. 2024. [Online] Available: https://jedfrees.github.io/Regression-Modeling-Online-Tutorial/

[32] Y. Yang, W. Qian, and H. Zou, "Insurance premium prediction via gradient tree-boosted Tweedie compound Poisson models," *Journal of Business & Economic Statistics*, vol. 36, no. 3, pp. 456–470, Jul. 2018, doi: 10.1080/07350015.2016.1200981.

[33] J. Pesantez-Narvaez, M. Guillen, and M. Alcañiz, "Predicting motor insurance claims using telematics data—XGBoost versus logistic regression," *Risks*, vol. 7, no. 2, Art. no. 70, Jun. 2019, doi: 10.3390/risks7020070.

[34] F. S. Khan, S. S. Mazhar, K. Mazhar, and M. A. Ashraf, "Model-agnostic explainable artificial intelligence methods in finance: a systematic review, recent developments, limitations, challenges and future directions," *Artificial Intelligence Review*, vol. 58, no. 8, Art. no. 232, Aug. 2025, doi: 10.1007/s10462-025-11215-9.

[35] R. Richman, "Applying deep learning in actuarial science," Social Science Research Network (SSRN), Rochester, NY, USA, SSRN Scholarly Paper 4749219, Dec. 2023, doi: 10.2139/ssrn.4749219.

[36] D. Chicco, M. J. Warrens, and G. Jurman, "The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation," *PeerJ Computer Science*, vol. 7, Art. no. e623, Jul. 2021, doi: 10.7717/peerj-cs.623.

[37] J. Mathew and N. Glidestone, "Explainable AI (XAI) in clinical decision support: Bridging transparency and trust in healthcare AI," *IEEE Journal of Biomedical and Health Informatics*, vol. 29, no. 6, pp. 2451–2462, Jun. 2025.

[38] N. S. Altman, "An introduction to kernel and nearest-neighbor nonparametric regression," *The American Statistician*, vol. 46, no. 3, pp. 175–185, Aug. 1992, doi: 10.1080/00031305.1992.10475879.

[39] 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, Dec. 2005, doi: 10.3354/cr030079.

[40] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed. New York, NY, USA: Springer, 2009.

[41] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, 1st ed. New York, NY, USA: Springer, 2013.

[42] B. Efron and R. J. Tibshirani, *An Introduction to the Bootstrap*. New York, NY, USA: Chapman & Hall/CRC, 1993, doi: 10.1201/9780429246593.

[43] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, Inc., 2017, pp. 4765–4774, doi: 10.48550/arXiv.1705.07874.

[44] M. T. Ribeiro, S. Singh, and C. Guestrin, ""Why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, Aug. 2016, pp. 1135–1144, doi: 10.1145/2939672.2939778.

[45] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning – A brief history, state-of-the-art and challenges," in *ECML PKDD 2020 Workshops*, C. Koprinska, M. Kamp, A. Appice, C. Loglisci, L. Antonie, A. Zimmermann, R. Gaudel, V. Lemaire, P. Ralha, and P. Ribeiro, Eds. Cham, Switzerland: Springer, 2021, pp. 417–431, doi: 10.1007/978-3-030-65965-3_28.