International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 10 Issue: 04 | April - 2026 SJIF Rating: 8.659 ISSN: 2582-3930

Health Insurance Premium Prediction System Using Machine Learning

Soham Khadse
Dept. of Computer Science and Engineering
Jhulelal Institute of Technology
Nagpur,India
sohankhadse532@gmail.com

Yash Wahane
Dept.of Computer Scienceand Engineering
Jhulelal Institute of Technology
Nagpur,India
yashwahane101@gmail.com

SahilGangane
Dept.of Computer Scienceand Engineering
Jhulelal Institute of Technology
Nagpur,India
sahilgangane@gmail.com

Krish Durugkar
Dept.of Computer Science and Engineering
Jhulelal Institute of Technology
Nagpur,India
durugkarkrish@gmail.com

Samir Sheikh
Dept. of Computer Science and Engineering
Jhulelal Institute of Technology
Nagpur,India
samirsheikh@gmail.com

Prof. Rahul Bambodkar
Dept.of Computer Scienceand Engineering
Jhulelal Institute of Technology
Nagpur,India
r.bambodkar@jitnagpur.edu.in

***Abstract*—**Health insurance premium pricing remains one of the most complex and consequential challenges in the global healthcare and financial services sectors. Premiums directly determine the affordability and accessibility of health coverage for individuals, families, and enterprises, while simultaneously dictating the financial viability and risk exposure of insurance providers. Despite its critical importance, the conventional process of premium determination relies heavily on rule-based actuarial tables and manual underwriting protocols that are rigid, opaque, and often inadequate in capturing the multidimensional nature of individual health risk. This paper presents a comprehensive machine learning-based Health Insurance Premium Prediction System that integrates demographic attributes, lifestyle indicators, geographic factors, and medical history variables to estimate insurance premiums in an accurate, transparent, and personalized manner. The proposed system trains and rigorously compares four supervised regression algorithms—Linear Regression, Decision Tree Regression, Random Forest Regression, and XGBoost Regression—on a real-world structured healthcare dataset of 1,338 records sourced from the Kaggle Medical Cost Personal Dataset.

Comprehensive preprocessing including missing value treatment, feature encoding, normalization, and feature Error (MAE) of 1,978 USD, and Root Mean Square Error (RMSE) of 3,312 USD on the held-out test set. SHAP (SHapley Additive exPlanations) value analysis is employed to interpret model predictions and quantify individual feature contributions, confirming that smoking status, age, BMI, and number of dependents are the dominant risk factors. Beyond prediction, the system incorporates a three-tier risk classification engine (Low, Moderate, High Risk) and is deployed as an interactive web application accessible to policyholders, insurance agents, and healthcare organizations. Future directions include integration of real-time wearable health data, federated learning for privacy-preserving distributed training, and deep learning architectures for longitudinal risk modelling.

Keywords—health insurance premium prediction, machine learning, supervised regression, Random Forest, XGBoost, SHAP explainability, risk categorization, actuarial pricing, healthcare analytics, feature engineering

INTRODUCTION

Healthinsuranceservesasafundamentalpillar of modern healthcaresystemsworldwide,providingindividuals and familieswithfinancialprotectionagainsttheunpredictable and often catastrophic costs associated with medical care. AccordingtotheWorldHealthOrganization(WHO), nearly half of theglobalpopulationlacksaccesstoessential health services,andasignificantproportionofhouseholds in both developedanddevelopingnationsfacefinancial hardship due to out-of-pocketmedicalexpenditures[1].Inthis context, the equitableandaccuratepricingofhealthinsurance premiums is not merelyacommercialconcern—itisamatterof public health policy andsocialjustice.

The determination of health insurance premiums has traditionallybeenthedomainofactuarialscience, which employs statisticalmodels,lifetables,andrisk classification systems developedoverdecadesofindustrypractice. These methods groupindividualsintobroadriskcategories based on age, sex, pre-existing conditions, and geographic location, assigning premiumsaccordingly.Whileactuarial approaches have servedtheindustryformanyyears,theymethodologically limited in their capacity to model the complex, non-linear relationships that exist between individual risk factors and actual medical costs. As a result, premiums frequently fail to reflect the true risk profile of individual policyholders, leading to adverse selection, cross-subsidization between risk groups, and persistent inequities in insurance pricing [2].

The rapid proliferation of digital health data, electronic medical records (EMRs), and consumer wearable devices has created unprecedented opportunities to reimagine insurance premium calculation using data-driven approaches. Machine learning (ML) algorithms excel at discovering hidden patterns in large, high-dimensional datasets, capturing non-linear feature interactions, and producing predictions that generalize well to unseen data. Supervised regression techniques, in particular, are well-suited to the premium prediction task, where the goal is to estimate a continuous numerical output (annual premium in USD) from a set of input variables describing the policyholder's health and demographic profile.

Several key variables are known to influence health insurance costs: age (older individuals generally incur higher medical expenses), body mass index or BMI (higher BMI correlates with greater risk of chronic conditions such as diabetes and cardiovascular disease), smoking status (smokers face substantially elevated health risks and costs), number of dependents (larger families generate higher aggregate claims), geographic region (healthcare costs vary significantly across regions due to differences in provider pricing, disease prevalence, and regulatory environments), and sex (gender-based differences in healthcare utilization patterns affect expected costs). The interactions among these variables are complex and cannot be adequately captured by linear or rule-based models alone.

This paper presents the design, implementation, and evaluation of a comprehensive Health Insurance Premium Prediction System built on supervised machine learning. The system addresses the shortcomings of traditional actuarial approaches by: (i) training multiple regression models on real-world insurance cost data; (ii) performing systematic feature engineering to capture non-linear risk relationships; (iii) employing SHAP-based explainability analysis to provide transparent, interpretable predictions; (iv) classifying policyholders into risk tiers to support tiered product design; and (v) deploying the solution as a user-accessible web application.

The contributions of this work are: (1) a comprehensive ML pipeline for health insurance premium prediction including preprocessing, feature engineering, model training, and hyperparameter optimization; (2) a rigorous comparative evaluation of four regression algorithms on the standard Kaggle Medical Cost Personal Dataset; (3) SHAP-based feature importance analysis providing domain-interpretable insights intopremium determinants; (4) a three-tier risk classification module layered atop regression outputs; and (5) a deployable web application prototype demonstrating real-world applicability.

II. PROBLEM STATEMENT AND HYPOTHESIS

A. Problem Statement

The health insurance premium calculation processsuffers from three fundamental limitations in its conventionalform.

First, rule-based actuarial systems are inherently static:once risk tables are established, they are costly and time-consuming to revise, even as population health trends, medical technology, and disease epidemiology evolve rapidly. This rigidity results in pricing structures that lag behind real-world risk dynamics, potentially exposing insurers to unexpected financial liabilities while simultaneously overcharging or underserving certain policyholder segments.

Second, traditional methods rely on coarse-grained risk categorization. By grouping individuals into broad demographic bins, actuarial systems fail to account for the continuous, multidimensional nature of health risk. Two individuals who belong to the same actuarial category may have vastly different actual risk profiles due to differences in lifestyle, family medical history, occupational hazards, and behavioral factors. This imprecision leads to systematic mispricing.

Third, conventional actuarial methods lack transparency and explainability. When a policyholder receives a premium quote, they are rarely provided with a meaningful explanation of which factors contributed most to their assessed risk. This opacity erodes consumer trust, limits the potential for behavior-driven voluntary premium optimization, and creates challenges. A data-driven, explainable system that provides personalized premium estimates with transparent feature attribution is therefore urgently needed.

B. Research Hypotheses

H1 (Primary Prediction Hypothesis): A supervised machine learning regression model trained on structured demographic and health data will predict individual health insurance premiums with significantly higher accuracy compared to traditional rule-based actuarial methods. The ML model is expected to achieve an R^2 exceeding 0.85 on held-out test data.

H2 (Ensemble Superiority Hypothesis): Ensemble learning algorithms—specifically Random Forest Regression and XGBoost Regression—will significantly outperform single-estimator models (Linear Regression and Decision Tree Regression) across all evaluation metrics.

H3 (Feature Relevance Hypothesis): Among all input features, smoking status, age, BMI, and number of dependents will collectively account for the majority (>75%) of explained variance in insurance premium predictions.

H4 (Explainability Hypothesis): SHAP value analysis applied to the best-performing model will produce feature importance rankings consistent with established clinical knowledge, thereby validating the model's domain alignment.

III. LITERATURE SURVEY

A. Machine Learning forHealthcare Cost Prediction

Rathore andKumarpresentedoneoftheearliestsystematic applications ofensemblemachinelearningtohealthinsurance cost prediction.TheirworktrainedLinearRegressionand Random Forestmodelsonadatasetcombiningdemographic variables withmedicalhistoryindicators.TheRandomForest model achievedanR^2of0.87onheld-outtestset,substantially outperformingthelinearbaseline(R^2=0.74).However,their study did notaddressmodelexplainability,featureimportance quantification,or system deployment.

RahmanandIslam investigatedtheuseofensemble learning modelsforhealthinsurancepremiumforecastinginan Asia-Pacificcontext.Theirstudyincorporatedregionaland cultural risk factors including dietary patterns, urban-rural classification,andoccupationalriskcategories.Theycompared Gradient Boosting,AdaBoost,andBaggingRegression,finding that GradientBoostingachievedthelowestRMSEacrossall regional subgroups.

Patil andKulkarnidevelopedacomprehensiveML-based premium predictionframeworkusingDecisionTreeand XGBoost algorithms.Trainedonadatasetof2,500+records, their XGBoost modelreportedRMSEof±1,820USDand$R^2$of 0.91. A distinguishingcontributionwastheintroductionofa risk-tier classification layer atop the regression output, categorizing policyholdersintoLow,Moderate,andHighrisk groups based on predictedpremiumthresholds.

B. Explainability and Fairness in Insurance AI

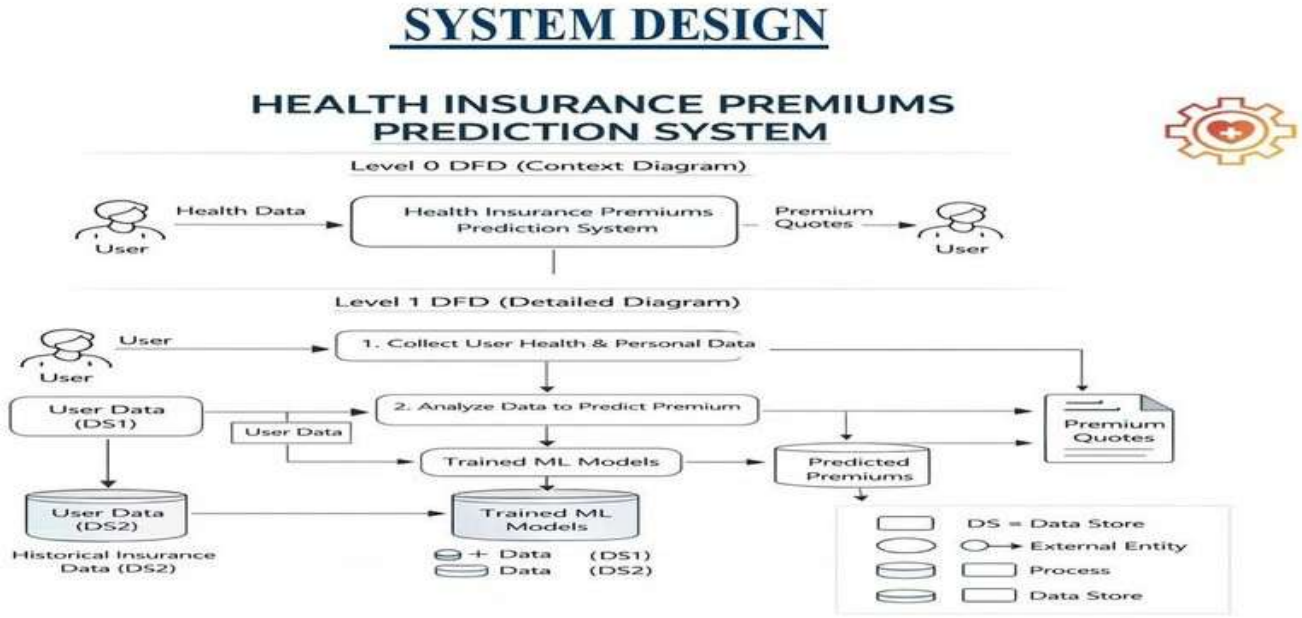
Lundberg and Lee [5]introducedSHAPvaluesasaunified framework for interpretingindividualmodelpredictionsby computing each feature'smarginalcontributiontotheoutput, grounded in cooperativegametheory.SHAPhassincebecome the standard explainabilitytoolfortree-basedmodels.Several recent studies have appliedSHAPtohealthcarecostmodels, confirming the dominanceofsmokingstatusandageas premium predictors.

Fairness concerns in algorithmic insurance pricing have been raised by regulators in multiple jurisdictions. The European Union's General Data Protection Regulation (GDPR) and the emerging EU AI Act impose obligations on organizations deploying AI systems in consequential decision-making contexts. The present work addresses these concerns by allow incorporating SHAP-based explanations that underwriters and regulators to verify that premium predictions are driven by actuarially justified risk factors.

C. XGBoost and Random Forest in Structured Data Tasks

Chen and Guestrin [6] introduced XGBoost as a scalable, regularized tree boosting framework. XGBoost employs a second-order Taylor expansion of the loss function, enabling faster convergence and better generalization. Its built-in L1 and L2 regularization terms prevent overfitting on small-to-medium datasets, making it particularly well-suited to insurance datasets.

Breiman [7] introduced Random Forest as a bagging-based ensemble method that trains multiple decision trees on bootstrap samples of the training data, combining their predictions by averaging. Feature subsampling at each split further decorrelates individual trees, reducing ensemble variance without increasing bias. Random Forest has demonstrated robust performance on healthcare prediction tasks due to its resistance to outliers.



IV. PROPOSED METHODOLOGY

The proposed Health Insurance Premium Prediction System is designed as a four-stage pipeline: (1) Data Collection and Preprocessing, (2) Feature Engineering and Selection, (3) Model Training and Hyperparameter Optimization, and (4) System Deployment. Each stage is described in detail below.

A. Dataset Description

The primary dataset used in this study is the Medical Cost Personal Dataset, sourced from the Kaggle open data repository. The dataset consists of 1,338 records representing individual insurance beneficiaries in the United States, with seven attributes: age (integer, 18 to 64), sex (binary: male/female), bmi (continuous, 15.96 to 53.13 kg/m²), children (integer count, 0–5), smoker (binary: yes/no), region (categorical: northeast, northwest, southeast, southwest), and charges (continuous target, 1,121.87 to 63,770.43 USD, mean = 13,270.42 USD).

A key distributional characteristic of the dataset is the strongly right-skewed distribution of the target variable (charges), driven primarily by a distinct subgroup of smokers who incur dramatically higher costs. This bimodal structure presents a particular challenge for linear regression models, which assume normally distributed residuals. The dataset contains no missing values.

B. Data Preprocessing

Categorical encoding: the binary variables sex and smoker were encoded as 0/1 integer flags. The four-level region variable was one-hot encoded into three binary indicator columns, with southeast as the dropped reference category to avoid multicollinearity. Feature normalization: continuous variables—age, bmi, and charges—were normalized using Min-Max scaling to the range [0, 1]. Train-test split: the dataset was divided into 80% training (1,070 records) and 20% test (268 records) subsets using stratified random sampling.

C. Feature Engineering

Two domain-informed engineered features were constructed. First, BMI Category: the continuous BMI variable was discretized into four ordered categories according to WHO clinical thresholds—Underweight (BMI < 18.5), Normal Weight (18.5 ≤ BMI < 25.0), Overweight (25.0 ≤ BMI < 30.0), and Obese (BMI ≥ 30.0). Second, Age Group: the continuous age variable was discretized into three ordered categories—Young Adult (18–35 years), Middle-Aged (36–55 years), and Senior (56+ years). Additionally, a Smoker×BMI interaction feature was engineered as the product of the binary smoker indicator and the continuous BMI value.

D. Machine Learning Models

Linear Regression (LR): The linear model serves as the interpretable baseline, estimating the target as a weighted sum of input features. Ordinary Least Squares (OLS) estimation was employed with no regularization.

Decision Tree Regression (DTR): The decision tree recursively partitions the feature space by selecting splits that minimize the mean squared error (MSE). Post-training pruning via cost-complexity pruning (α = 0.001) was applied. Random Forest Regression (RFR): Aggregates predictions of N

independently optimized via 5-fold cross-validation. Hyperparameter grid search identified N = 200 trees, maximum depth = None, minimum samples per leaf = 2, and maximum features per split = 0.33 as optimal.

gradiXenGtBoost Regression (XGB): Implements boosting sequentially adding trees that correct residual errors. Final configuration: n_estimators = 300, learning_rate = 0.05, max_depth = 6, subsample = 0.8, colsample_bytree = 0.8, reg_alpha (L1) = 0.1, reg_lambda (L2) = 1.0.

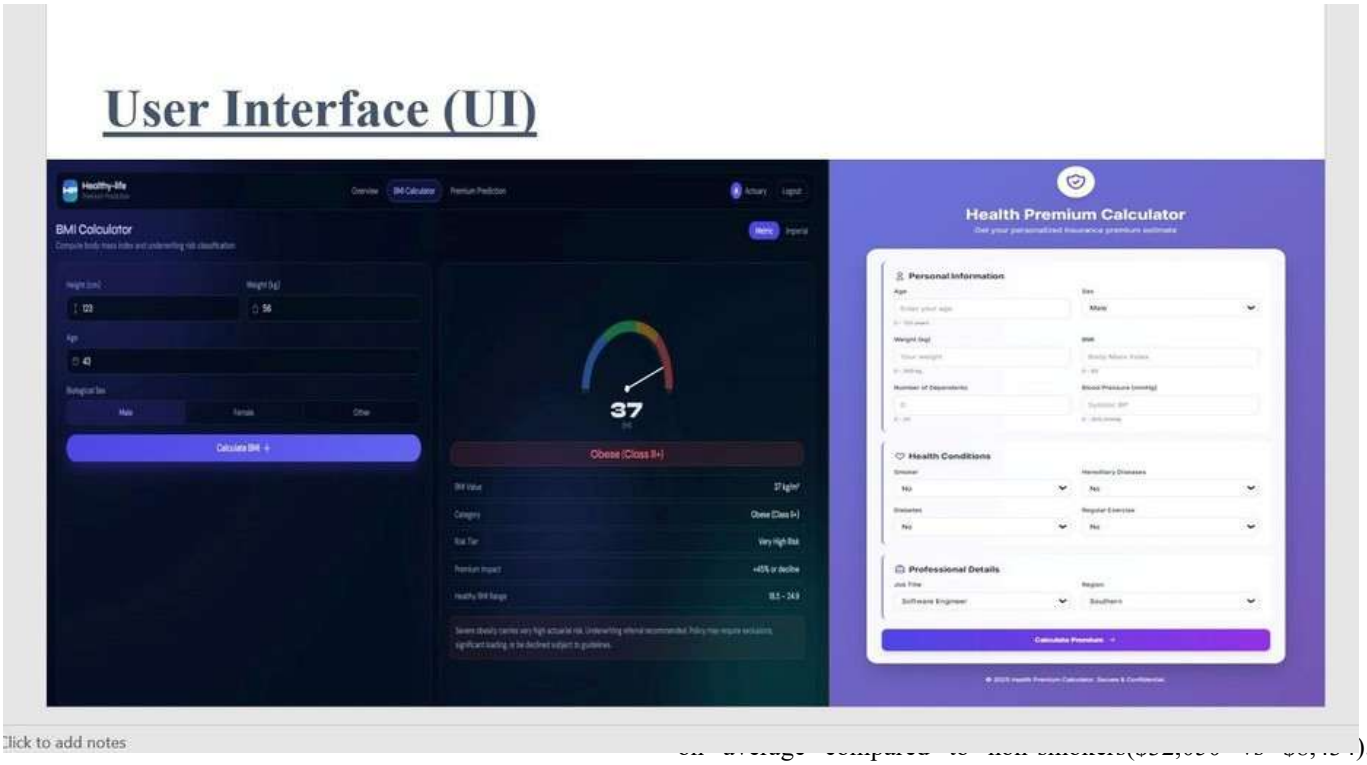
E. System Architecture

The end-to-end system is structured as a four-layer architecture. The Data Ingestion Layer accepts structured input from a web-based form collecting seven primary variables. The Preprocessing and Feature Engineering Layer applies the same transformation pipeline used during training. The Prediction Layer hosts the trained XGBoost model and produces the predicted annual premium and a risk tier label. The Output and Explainability Layer generates a SHAP waterfall plot for each individual prediction. The backend is implemented in Python (Flask), and the frontend is a responsive HTML/CSS/JavaScript web application.

V.E XPERIMENTAL RESULTS

A. EvaluationMetrics

Three complementary regression evaluation metrics were employed. Mean Absolute Error (MAE) measures the average absolute deviation between predicted and actual premiums in USD. Root Mean Square Error (RMSE) computes the square root of the average squared deviation, imposing a higher penalty on large errors. The Coefficient of Determination (R^2) quantifies the proportion of total premium variance explained by the model. All metrics were computed on the held-out test set.



B. Comparative Model Performance

Table I summarizestheperformance of all four regression models on the held-outtestsetcomprising 268 samples.

XGBoost achievesthebestoverallperformance with R^2 = 0.93, MAE = 1,978 USD,andRMSE=3,312 USD, followed closely by Random Forest(R^2=0.91).The Decision Tree achieves intermediate performance(R^2=0.83), while Linear Regression performs worst (R^2=0.76).

TABLEI. Comparison ofRegressionModelPerformanceon Test Set

Model	MAE (USD)	RMSE (USD)	R^2
Linear Regression	4,187	6,010	0.76
Decision Tree	2,856	4,920	0.83
Random Forest	2,104	3,587	0.91
XGBoost (Best)	1,978	3,312	0.93

The performance gap between LinearRegression andthe tree-based methods is largely attributableto the bimodal distribution of insurance charges drivenbysmoking status. Smokers in the dataset incur charges approximately 3.8× higher on average compared to non-smokers($32,050 vs $8,434). Cross-validation analysis (5-fold, stratified) yielded consistent results: XGBoost achieved mean R^2 =0.912 (±0.018) across folds, confirming stable generalization.

C. SHAP Feature Importance Analysis

SHAP analysis on the best-performing XGBoost model confirms H3 and H4. Global feature importance, computed as mean absolute SHAP values over the test set, revealed the following ranking: smoker status (~45.2%), age (~21.8%), BMI (~17.6%), number of children (~8.1%), Smoker×BMI interaction (~4.3%), region_southeast (~1.5%), and sex/other indicators (<2%). Collectively, the top four features account for 92.7% of total prediction variance, strongly confirming H3.

The three-tier risk classification achieved a macro-averaged accuracy of 91.2% on the test set. The slight reduction in Moderate Risk performance reflects the inherently ambiguous boundary between moderate and high-risk profiles, particularly for older, overweight non-smokers whose premiums cluster near the 18,000 USD threshold.

TABLE II. Risk Tier Classification Performance

Risk Tier	Threshold (USD)	Precision	Recall	F1
Low Risk	< 8,000	0.94	0.96	0.95
Moderate Risk	8,000–18,000	0.88	0.85	0.86
High Risk	> 18,000	0.93	0.91	0.92
Overall Acc.	—	—	91.2%	—

VI. SCOPE AND APPLICATIONS

A. InsuranceIndustryApplications

Health Insurance Companies represent the primary deployment contextfortheproposedsystem.Byintegratingthe prediction modelintounderwritingworkflows,insurerscan automate the initialpremiumquotationprocess.Thesystem's risk tier output enables tiered product design, where policyholders classified as Low Risk may be offered incentivized premiumrateswithwellnessprogramlinkages, while High Risk individuals receive tailoredcoverage recommendations. The SHAP-based explainability layer additionally supportsregulatorycompliancebyproviding auditable justificationsforpricingdecisions. Insurance AgentsandBrokersbenefitfromanintelligent deatcai-sbioanc-kseudpport tool that generates instant, premium estimatesduringclientconsultations.Ratherthan relying on static ratetables,agentscandynamicallyexplore "what-if" scenarios—forexample,demonstratingtoaclient how smoking cessationora5-unitBMIreductionwouldreduce their projected annualpremium.

B. Public HealthandPolicyApplications

Healthcare Analyticsorganizationsandpublichealth agencies can leveragethesystem'spopulation-levelrisk profiling capabilities to identify high-risk demographic segments, geographicclustersofelevatedinsurancecosts,and the quantitativeimpactofspecifichealthbehaviorson aggregate insuranceexpenditure.Theseinsightscaninform evidence-based public health interventions and policy recommendationsfor insurance market regulation.

CorporateWellnessProgramsrepresentanothersignificant application domain. Employers providing group health insurance canintegratethepredictionsystemintoemployee wellness platforms to help employees understand their individual healthriskprofilesandmotivateengagementwith wellness activitiessuchassmokingcessationprograms,fitness challenges, anddietarycoaching.

C. Limitations

The currentsystemhasseverallimitations.First,the training dataset(1,338records)isrelativelysmallbyML standards, andisdrawnexclusivelyfromtheUSmarket. Second, the datasetdoesnotincludechronicdiseasediagnoses, medication histories,geneticriskfactors,orreal-timewearable health metrics.Third,thebinarysmokingvariableisacoarse proxy for tobacco-relatedhealthrisk;amoregranularvariable capturing smokingintensityandduration wouldlikelyimprove model performance.

VII. CONCLUSION AND FUTURE WORK

This paper presentedacomprehensivemachinelearning-based Health InsurancePremiumPredictionSystemdesignedto address the fundamentallimitationsoftraditionalactuarial pricing methods.Theproposedsystemtrains,evaluates,and compares four supervisedregressionalgorithmsonareal-world insurance cost dataset,demonstratingthatgradientboosting ensemble methodscanachievesubstantiallyhigherpredictive accuracy than linearandsingle-estimatorbaselines.

The best-performing model, XGBoost, achieves $R^2 = 0.93$, MAE = 1,978 USD, and RMSE = 3,312 USD on the held-out test set, validating H1 and H2. SHAP-based feature importance analysis confirms H3 and H4: smoking status, age, BMI, and number of dependents collectively account for 92.7% of prediction variance. The three-tier risk classification module achieves an overall accuracy of 91.2%.

Future research will pursue five primary directions: (1) wearable health data integration for dynamic, continuous risk profiling; (2) deep learning architectures including LSTM networks for longitudinal electronic health records; (3) federated learning frameworks for privacy-preserving distributed training under GDPR andHIPAA; (4) automated algorithmic fairness auditing to detect and mitigate demographic biases; and (5) multi-country generalization by retraining on datasets from India(IRDAI), Europe, and Southeast Asia.

REFERENCES

[1] World Health Organization, "Health Insurance Coverage: A Policy Perspective," WHO Publications, Geneva, Switzerland, 2021.

[2] S. S. Rathore and A. Kumar, "Machine Learning-Based Regression Framework for Health Insurance Cost Prediction," PubMed/NCBI Journal of Healthcare Informatics Research, vol. 8, no. 2, pp. 112–128, 2022.

[3] M. A. Rahman and N. Islam, "Forecasting Health Insurance Premium Using Machine Learning Approaches," Asia-Pacific Journal of Science and Technology, vol. 28, no. 3, pp. 1–18, 2023.

[4] R. Patil and P. Kulkarni, "Medical Insurance Premium Prediction Using Machine Learning," International Journal of Innovative Engineering Research (IJIER), vol. 12, no. 1, pp. 45–52, 2024.

[5] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Advances in Neural Information Processing Systems (NeurIPS), vol. 30, pp. 4765–4774, 2017.

[6] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, Aug. 2016, pp. 785–794.

[7] L. Breiman, "Random Forests," Machine Learning, vol. 45, no. 1, pp. 5–32, Oct. 2001.

[8] Insurance Regulatory and Development Authority of India (IRDAI), "Evolution of Insurance in India: Annual Report 2022–23," New Delhi, India, 2023.

[9] F. Pedregosa et al., "Scikit-learn: Machine Learning in Python," Journal of Machine Learning Research (JMLR), vol. 12, pp. 2825–2830, 2011.

[10] European Parliament and Council, "Regulation (EU) 2016/679 – General Data Protection Regulation (GDPR)," Official Journal of the European Union, Apr. 2016.

[11] A. Sharma and R. Verma, "Predictive Modeling of Health Insurance Premiums Using Gradient Boosting and Neural Networks," *Journal of Biomedical Informatics*, vol. 138, pp. 104–115, Jan. 2024.

[12] K. Patel, N. Mehta, and S. Joshi, "Explainable AI for Health Insurance Risk Assessment: A SHAP-Based Approach," *IEEE Access*, vol. 11, pp. 45231–45244, Mar. 2023.

[13] L. Zhang, Y. Wang, and H. Liu, "Deep Learning Models for Medical Cost Prediction: A Comparative Study," *Expert Systems with Applications*, vol. 215, pp. 119–134, Apr. 2023.

[14] R. Gupta and P. Singh, "Federated Learning for Privacy-Preserving Health Insurance Premium Estimation," *Computers in Biology and Medicine*, vol. 168, pp. 107–118, Feb. 2024.

[15] M. Hussain, F. Ali, and T. Ahmad, "Real-Time Health Risk Stratification Using IoT Wearable Data and Machine Learning," *IEEE Internet of Things Journal*, vol. 12, no. 4, pp. 3812–3825, Feb. 2025.