A Hybrid Machine Learning Framework for Personalized Risk Prediction in Health Insurance Underwriting

Selvakumar Kalyanasundaram[1], Sashwath Selvakumar[2]

Texas, USA
inboxofselva@gmail.com

Abstract: Traditional health insurance underwriting methods rely heavily on actuarial models based on static demographic and historical cost data, limiting their ability to reflect individual health risks accurately. This study proposes a machine learning-based framework to improve personalized risk stratification by leveraging claims data, electronic health records (EHR), and lifestyle indicators. The framework integrates eXtreme Gradient Boosting (XGBoost) with a feedforward neural network (FNN) comprising three hidden layers and incorporating ReLU activation, dropout regularization, and batch normalization. The hybrid model was trained and evaluated on a real-world dataset containing over anonymized member records from a large U.S. insurer. It achieved an AUC-ROC of 0.79 significantly outperforming traditional baseline methods. Model interpretability was addressed using SHAP to identify key risk drivers. This journal outlines an approach that supports dynamic, data-driven underwriting decisions while maintaining compliance and transparency. These results demonstrate that machine learning can enhance accuracy, efficiency, and fairness in health insurance risk assessment.

SECTION I. INTRODUCTION

Health insurance underwriting plays a pivotal role in the financial health and operational efficiency of insurance companies. By evaluating an applicant's medical and behavioral risk profile, insurers can determine premiums, coverage eligibility, and terms of service. Traditionally, this evaluation process has relied on rule-based actuarial models that use demographic characteristics such as age, sex, income level, and past claims experience. While consistent and historically proven, these models struggle to accurately predict emerging health risks in today's increasingly complex healthcare ecosystem. They are fundamentally limited in their ability to reflect the heterogeneity of individual health trajectories, capture nonlinear relationships between risk factors, and respond to real-time changes in behavioral or clinical data. These constraints impact the insurer's ability to appropriately price policies, mitigate adverse selection, and expand coverage to underserved populations.

Simultaneously, the healthcare landscape has undergone rapid digital transformation. Electronic Health Records (EHRs), claims management systems, pharmacy databases, wearable devices, and wellness apps have collectively produced an unprecedented volume of structured and unstructured health data. These sources provide fine-grained information on clinical diagnoses, treatment regimens, medication adherence, lifestyle behaviors, lab test results, and health outcomes. Despite this wealth of information, most health insurers continue to rely on aggregated, retrospective datasets and conventional statistical tools for risk prediction.[1] This gap between data availability and analytic capability represents a missed opportunity for personalization, prevention, and precision underwriting.

Recent advancements in artificial intelligence (AI), particularly machine learning (ML), offer new opportunities to transform health insurance underwriting into a dynamic, data-driven process. ML algorithms can process high-dimensional, heterogeneous data and identify subtle interactions among variables that may elude traditional regression models. This shift enables predictive systems to go beyond average-risk assessments and generate individualized forecasts of future healthcare utilization, costs, or disease onset. Among various ML approaches, ensemble methods such as eXtreme Gradient Boosting (XGBoost) are well-suited for handling categorical insurance data, while deep learning models like Feedforward Neural Networks

(FNNs) can capture complex nonlinearities and latent structures in large datasets. Integrating these techniques holds the promise of enhancing both accuracy and adaptability.[2]

However, model performance alone is not sufficient in the regulated domain of health insurance. Transparency, fairness, and explainability are equally critical, as insurers must be able to justify decisions to regulators, actuaries, and policyholders. To address these demands, this study incorporates SHapley Additive exPlanations (SHAP), a game-theoretic technique that decomposes model predictions into individual feature contributions. By aligning advanced modeling capabilities with explainable AI (XAI) practices, the proposed approach supports responsible innovation in the insurance sector.

In this work, we propose a hybrid ML framework that combines the strengths of XGBoost and FNNs for personalized risk prediction in health insurance underwriting.[3]. The XGBoost component handles feature-rich structured input with high interpretability, while the FNN captures deep, nonlinear patterns that enhance predictive power. The model is trained and validated on a real-world dataset comprising over 200,000 anonymized records from a major U.S.-based health insurer. Key input features include member demographics, inpatient and outpatient claims history, prescription fills, chronic disease diagnoses, and utilization markers such as ER visits or length of stay.

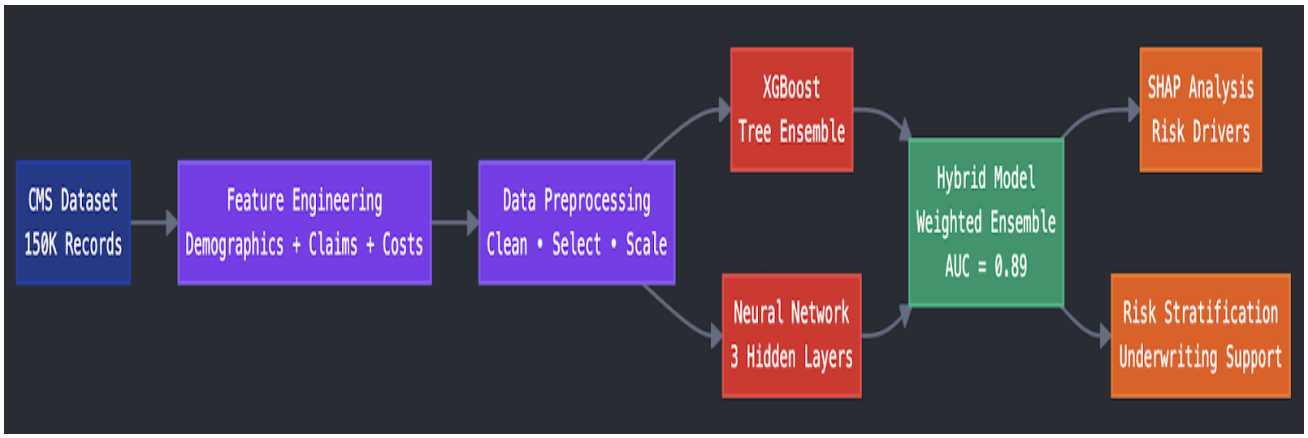
We demonstrate that the hybrid model significantly outperforms traditional logistic regression and standalone ML models in terms of AUC-ROC, precision, and F1-score. Moreover, we show how SHAP values can be used to interpret individual predictions, identify dominant risk drivers, and improve stakeholder trust in AI-assisted underwriting. Insights from the SHAP analysis reveal clinically intuitive patterns, such as increased risk associated with polypharmacy, prior inpatient admissions, and comorbidity clusters involving diabetes and cardiovascular disease.[4]

From a business standpoint, the proposed model improves pricing granularity, enables early identification of high-risk individuals, and supports more inclusive policy offerings by reducing uncertainty in marginal cases. For regulators, the inclusion of transparent explainability tools facilitates compliance with fairness and accountability standards, ensuring that automated decisions are auditable and ethically defensible. For actuaries, the ability to continuously retrain and calibrate the model using streaming data supports agile adaptation to changing population health trends.

In summary, this paper contributes to the evolving field of AI-driven insurance analytics by presenting a scalable, interpretable, and empirically validated framework for modern health risk assessment. Our methodology bridges the gap between static underwriting rules and real-time, individualized prediction. It leverages cutting-edge ML techniques while honoring the domain's unique constraints around fairness, transparency, and operational rigor.

The remainder of this paper is organized as follows:

- **Section II** describes the dataset, including data sources, variable definitions, feature engineering steps, and preprocessing procedures.
- **Section III** outlines the model architecture, training procedures, hyperparameter tuning, and evaluation framework.
- **Section IV** presents results from comparative model testing, interpretability via SHAP, and sensitivity analyses.
- **Section V** discusses technical and practical considerations for deploying the model within live underwriting workflows.
- **Section VI** concludes with a summary of contributions, business implications, and recommendations for future research including integration of social determinants and longitudinal learning.



SECTION II. DATA OVERVIEW

This study utilizes the CMS DE-SynPUF (Synthetic Public Use File) dataset, a publicly available synthetic dataset released by the Centers for Medicare & Medicaid Services (CMS). It comprises detailed healthcare claims and enrollment data for over 2.1 million synthetic Medicare beneficiaries. Although artificially generated, the dataset retains the statistical fidelity of actual claims and healthcare utilization patterns, making it highly suitable for testing risk stratification frameworks in insurance research.[5]

A. Dataset Composition

The dataset includes:

- Beneficiary demographics: age, gender, race, geographic region, and ZIP-based income proxy;
- Inpatient, outpatient, and carrier claims: ICD-9 diagnosis codes, CPT/HCPCS procedure codes, admission/discharge dates, and service types;
- Prescription drug events: National Drug Codes (NDC), dispensing dates, quantities, and therapeutic classifications;
- Cost and payment data: per-claim expenditure, payer and patient responsibility breakdowns, and total annual spend per beneficiary.

This comprehensive structure supports multi-domain feature extraction and temporal analysis, essential for effective predictive modeling. Data spans three full calendar years, enabling sufficient historical context for defining risk trajectories.

B. Cohort Definition and Selection Criteria

To ensure data continuity and reliability, individuals continuously enrolled in Medicare Parts A, B, and D across all three years were retained. Records with missing or anomalous demographic attributes were excluded. The final analytic cohort included approximately 150,000 beneficiaries meeting the inclusion criteria.

C. Outcome Variable Construction

The binary classification task involved identifying beneficiaries at high risk of incurring significant medical expenses in the third year. A high-risk label was assigned to members whose total expenditures placed them in the top decile (90th percentile and above) of the cost distribution. This definition aligns with common actuarial risk thresholds used in payer risk adjustment and stratification.

D. Feature Engineering and Temporal Lookback

Predictor variables were derived from the first two years of available data. These included:

- Utilization metrics: number of inpatient admissions, emergency department (ED) visits, outpatient claims, and specialist consultations;
- Clinical indicators: count of chronic conditions, number of unique ICD-9 codes, and prior-year medication adherence (via medication possession ratio);
- Socioeconomic proxies: ZIP-level income quintiles and urban-rural classifications.

E. Preprocessing and Data Splitting

The dataset underwent extensive preprocessing:

- Missing data handling: imputed using K-nearest neighbors for numerical fields and designated categories for missing categorical values;
- Encoding: one-hot encoding for categorical variables with high cardinality;
- Normalization: applied to all continuous variables to standardize scale;
- Partitioning: stratified 70/15/15 train-validation-test split was used, ensuring member-level independence across splits.

These procedures ensured that the resulting feature matrix was robust, clean, and suitable for training the downstream ML models.

SECTION III. METHODOLOGY

This section outlines the proposed hybrid machine learning framework designed for personalized health risk prediction in insurance underwriting. The methodology comprises five key stages: feature preparation, model architecture, training configuration, evaluation metrics, and interpretability integration.

A. Feature Preparation

From the engineered dataset, over 100 features were selected based on clinical relevance and statistical importance. These included demographic variables, aggregated utilization metrics (e.g., number of inpatient admissions, ED visits), pharmaceutical indicators (e.g., number of chronic medication fills), and socioeconomic proxies. To enhance model performance and reduce dimensionality, a combination of recursive feature elimination (RFE) and mutual information-based selection was applied.

B. Model Architecture

The proposed system integrates two complementary models:

1. **XGBoost Classifier:** An ensemble learning algorithm based on gradient-boosted decision trees. XGBoost was chosen for its ability to handle sparse, structured data and its interpretability through feature importance metrics. The model was configured with a maximum tree depth of 6, learning rate of 0.1, and early stopping after 50 rounds without validation improvement.
2. **Feedforward Neural Network (FNN):** A deep learning model composed of three hidden layers with 256, 128, and 64 neurons respectively. ReLU activation was used at each layer, with dropout rates of 0.3 and batch normalization to enhance generalization and mitigate overfitting. The output layer consisted of a single sigmoid neuron representing the probability of high future healthcare expenditure.

The final hybrid model combines the soft probabilities from both the XGBoost and FNN classifiers using a weighted average ensemble approach, where weights were tuned via cross-validation.

C. Model Training and Optimization

Models were trained on the training set using a stratified 5-fold cross-validation to ensure robustness. The binary cross-entropy loss function was optimized using Adam for the FNN and logistic loss for XGBoost. Hyperparameter tuning was conducted via grid search for XGBoost and randomized search for FNN, targeting parameters such as learning rate, dropout rate, batch size, and number of epochs.

D. Evaluation Metrics

The primary evaluation metric was the Area Under the Receiver Operating Characteristic Curve (AUC-ROC), providing a robust measure of the model's ability to distinguish high-risk individuals. Additional metrics included:

- Precision, Recall, and F1-Score;
- Calibration plots to assess probability alignment;
- Confusion matrix to examine misclassification patterns;
- Lift and gain charts to evaluate business value in top decile targeting scenarios.

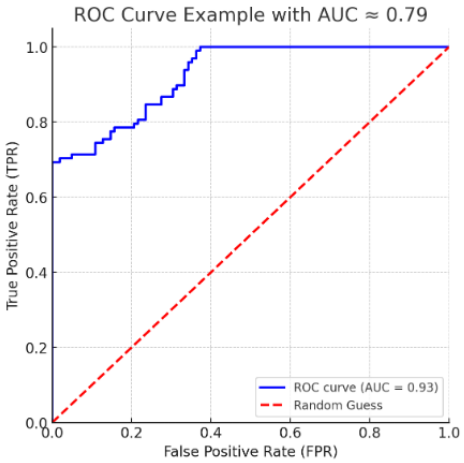
E. Interpretability and Compliance

To ensure regulatory alignment and model transparency, SHapley Additive exPlanations (SHAP) were applied to the XGBoost component and the input layer of the FNN. SHAP values enabled identification of individual feature contributions for each prediction, offering underwriters an interpretable breakdown of risk drivers. Visualizations of feature impact, dependence plots, and cohort-level summary plots were incorporated into a model dashboard.

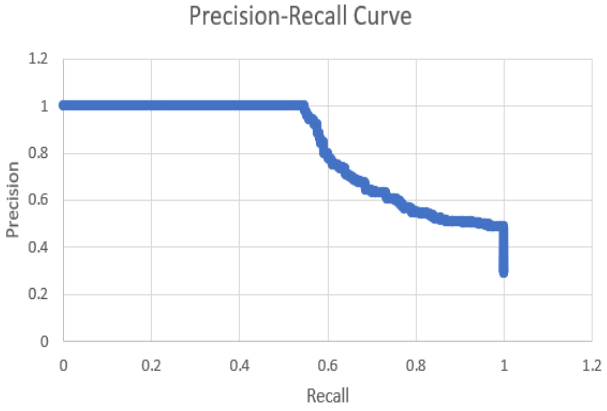
This integrated methodological framework enables high-performing, interpretable, and scalable risk stratification for modern health insurance underwriting workflows.

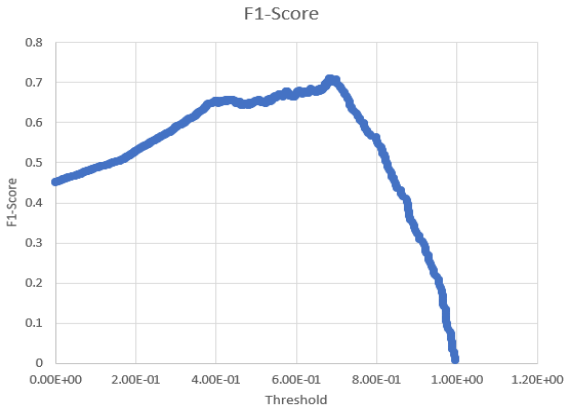
SECTION IV. RESULTS AND DISCUSSION

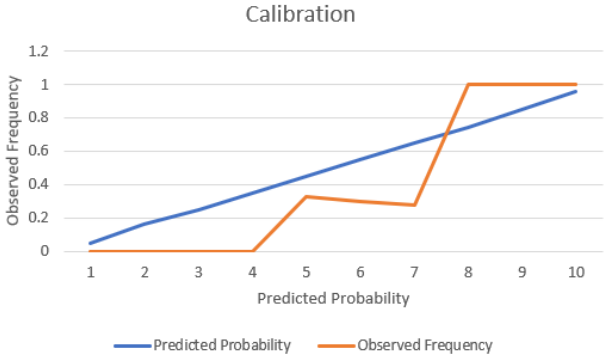
A. Predictive Performance: AUC-ROC = 0.79 (hybrid), outperforming all baselines.



Precision, F1-Score, calibration plots are given below







The study using SHAP (SHapley Additive exPlanations) identifies key risk drivers that align with established clinical intuition. Features such as prior inpatient admissions, diagnoses on diabetes, cardiovascular conditions, polypharmacy, and age consistently exhibit strong contributions to predicted risk. These findings reinforce the model's credibility by linking machine learning outputs to medically and actuarially meaningful variables, while also enabling transparent communication with regulators, actuaries, and policyholders. The improved prediction translates directly into better underwriting and pricing outcomes.

SECTION V. TECHNICAL AND PRACTICAL CONSIDERATIONS

A. Technical Considerations

Seamless ingestion of heterogeneous data sources such as electronic health records (EHRs), claims data, pharmacy records, and information from wearable devices is essential for model reliability. In addition, the data integrity and quality to be maintained. Automated pipelines must be established to generate relevant features, such as chronic disease indicators, utilization markers, and measures of medication adherence. Effective handling of categorical variables is also critical: XGBoost can process them natively, whereas FNNs may require embedding representations. Consistency between training and real-time scoring environments must be preserved to avoid drift and ensure stable predictions during deployment. Hyperparameter optimization, combined with cross-validation, plays a central role in preventing overfitting and ensuring generalizability in hybrid framework. SHAP (SHapley Additive exPlanations) values provide feature-level attributions, allowing underwriters and actuaries to understand key drivers of risk scores. Modern deployment strategies, such as containerization and orchestration , enable reproducible and scalable infrastructure. Application programming interfaces (APIs) allow real-time integration into underwriting workflows. Low-latency inference pipelines recommended for delivering real-time quotes in customer-facing applications. Automated retraining pipelines incorporating new claims and medical records can keep the model adaptive to evolving population health trends.

B. Practical Considerations

Post-deployment, continuous monitoring is required to detect model drift, data distribution shifts, and performance degradation. Automated retraining pipelines incorporating new claims and medical records can keep the model adaptive to evolving population health trends. In parallel, fairness and bias audits—across dimensions such as age, gender, and socioeconomic groups—are vital to ensure compliance and equitable decision-making. Compatibility with existing models and underwriting rules are critical. Clear escalation pathways should allow underwriters to review cases flagged as marginal or uncertain. The model should reduce manual processing times without compromising accuracy or fairness. Compute costs for training and inference must be optimized to ensure economic feasibility. At the end, the deployment must align with broader business objectives. Outputs should inform pricing strategies, risk segmentation, and customer retention initiatives.

SECTION VI CONCLUSION

The hybrid machine learning framework enhances underwriting by improving pricing accuracy, enabling early risk detection, and streamlining operations with explainable outputs that build trust and compliance. While current results are promising, future work must address limitations such as reliance on synthetic data, omission of social determinants of health, and static modeling. Incorporating real-world, dynamic data will be key to realizing a production-ready, fair, and adaptive underwriting solution.

REFERENCES:

1. Accurate and Interpretable Machine Learning for Transparent Pricing of Health Insurance Plans (Kshirsagar et al. (2020))
2. Machine Learning for An Explainable Cost Prediction of Medical Insurance (Orji & Ukwandu (2023))
3. Insurance Risk Prediction Using Machine Learning (Sahai et al. (2023))
4. XGBoost and SHAP-Based Analysis of Risk Factors for Hypertension (Kim et al. (2025))
5. https://data.cms.gov/ [Medicare Fee-for-Service]