Journal of Trends in Computer Science and Smart Technology (ISSN: 2582-4104)
www.irojournals.com/tcsst/

Deployment-Oriented Evaluation of Healthcare Reimbursement Cost Prediction

Hayat Ouadi[1], Ilhame El Farissi[2], Ilham Slimani[3]

Mohammed First University Oujda, Morocco.

E-mail: [1]hayat.ouadi@ump.ac.ma, [2]i.elfarissi@ump.ac.ma, [3]slimani.ilham@gmail.com

Abstract

The use of healthcare cost prediction models to assist with resource allocation and risk management is on the rise, but their use is limited by the requirement for reliable, easily interpretable explanations. Although there are multiple methods of providing these explanations, their operational use in healthcare settings remains inadequately evaluated. This research analyses four of the available explainable models (i.e., Permutation Feature Importance, Tree Importance, Local Interpretable Model Explanations, and Shapley-based explanations) using a Random Forest healthcare cost prediction model developed on a real-world dataset (comprising 2,302 aggregated patient segments and 54 features) with a coefficient of determination of 0.9957. The analysis used three criteria that are critical to the deployment of these models: steadiness of explanation under data perturbation, correspondence of feature importance across methods, and computational latency. Results show that each of the global explainable models has relatively high steadiness and relatively low latency, making them applicable for real time and regulatory use; however, the local explainable models provide more intuitive instance-level explanations at the expense of lower steadiness and greater computational demands. A tiered deployment framework for method selection is proposed based on these results as guidance for selecting methods based on clinical, regulatory, and operational needs. This research provides insight into the practical use of explainable healthcare cost prediction systems within real-world environments.

Keywords: Explainable AI, Healthcare Cost Prediction, Feature Importance Analysis, Deployment Feasibility, Stability Interpretability Trade-Off.

1. Introduction

1.1 Healthcare Cost Prediction and the Regulatory Imperative

Globally, healthcare expenditure is an important issue for policymakers and the economy at large due to the combination of increasing numbers of older people, more individuals with chronic conditions, and the rising costs of medicines and hospital care[1]. In France, for example, total health expenditures are in the hundreds of billions of euros each year, which places enormous demands on public finances and insurance companies[2]. Therefore, accurate prediction of costs in healthcare is critical for many operational management activities such as budgeting, resource allocation, risk stratification, and detecting fraud. The recent trend toward using machine learning models for these types of tasks has occurred because they are better able to predict events in a wide variety of domains compared with traditional statistical and actuarial approaches.

DOI: https://doi.org/10.36548/jtcsst.2026.2.008
Received: 27.02.2026, received in revised form: 30.03.2026, accepted: 17.04.2026, published: 29.04.2026

Although machine learning models have been shown to be effective, they are often viewed as lacking transparency[3]. This has raised significant concerns, especially when it comes to making decisions regarding healthcare that impact people's lives. The concern around transparency in the use of machine learning for assisting in the making of healthcare decisions has been amplified by new regulations, such as the European Union's General Data Protection Regulation [4] (GDPR), which provides individuals with the right to receive an explanation for algorithmic-based decisions made about them, as well as other regulatory guidance that emphasizes the need for transparency and robustness in the use of medical artificial intelligence systems. In addition, the demand for transparency in algorithmic decision-making is being driven by health care professionals, who need to know whether or not the output of a machine learning model is clinically relevant, ethically acceptable, and free from systemic bias before they act on its recommendation[5]. Therefore, explainable artificial intelligence (XAI) will no longer be viewed as an option for improving predictive models in relation to predicting healthcare costs; instead it will be considered a minimum requirement.

1.2 Explainable Methods for Healthcare Models

A limited number of methods for providing effective explanations have been developed and proposed to support transparency issues arising from machine learning models, each of which is representative in relation to their application within healthcare. There are four main approaches currently being used as methods for providing model-independent explanations; Permutation Feature Importance (PFI)[6], Tree-based Feature Importance (TFI)[7], Local Interpretable Model Explanations (LIME)[8], and Shapley-Based Feature Importance (SFI)[9].Each of the four methods provides either a global or local explanations, and each offers different types of advantages and disadvantages, including but not limited to stability, interpretability, and computational cost.

While a significant amount of research has been considering using explainable AI (XAI) technologies, prior works often assess XAI technologies independently from one another and frequently focus purely on qualitative interpretability without considering relevant operational constraints. For instance, work by Hettikankanamage et al. (2025) [10] on XAI for biomedical imaging and Brandsæter & Glad (2025) [9] on Shapley values for prediction accuracy underscores a fundamental tension: explanation methods that perform well under ideal laboratory conditions often degrade in operational healthcare environments. Specifically, three critical gaps have emerged in the literature:

- **Stability Under Data Perturbation:** While theoretical stability has been established in isolated settings, a notable research gap remains regarding head-to-head comparisons of explanation consistency under stochastic stress. Traditional evaluations often assume pristine data environments, thereby failing to account for the performance degradation caused by noise, measurement inaccuracies, and the temporal drift inherent to clinical practice[5].

- **Cross-Method Consensus and Agreement:** Recent comparative reviews reveal that explainability methods diverge substantially beyond top-K features, yet there is a lack of research quantifying the level of agreement between different XAI technologies and its impact on clinical decision-making. High agreement on dominant features masks potential disagreement on secondary variables, which may prove clinically relevant in subset populations[11].

- **Computational Feasibility at Scale:** Latency (i.e., time taken) required by different XAI technologies when generating an explanation is absent from the healthcare XAI literature despite pervasive real-time requirements in operational settings. This omission limits the practical deployment of computationally intensive methods like SHAP and LIME [12].

1.3 Knowledge Gap, Proposed Approach, and Research Contribution

While there has been a significant amount of research done using explainable AI technologies, prior works often assess explainable AI technologies independently from one another and often focus purely on qualitative interpretability without taking into account relevant operational constraints. There remains a significant gap in the field concerning the systematic comparison of dimensions most relevant for deploying explainable AI technologies in real-world healthcare settings. These include; explanation stability (i.e., consistency) under data perturbation, the level of agreement between different explainable AI technologies, and the latency (i.e., time taken) required by different explainable AI technologies when generating an explanation [10]. This gap has significant ramifications when it comes time for healthcare organizations to choose between various explainable AI technologies that align with their regulatory, clinical and technical needs in terms of the use of explainable AI technologies [5].

To address this gap, this paper presents a comparative evaluation framework for four explainable AI technologies applied to the prediction of healthcare expenditures. All four methods were assessed on the basis of three critical operational use-case criteria; stability, alignment and latency.

An accompanying tiered deployment strategy has also been developed based on the empirical results obtained from this study, which provides recommendations for the appropriate selection of an explainable AI technology depending on the intended use, including real-time clinical alerts, regulatory compliance auditing, and exploratory data analysis. Thus, the primary contribution of this study is closing the gap between explainable AI theory and practical application in healthcare settings and providing quantifiable and actionable recommendations for the integration of explainable AI technologies into the healthcare system for the purpose of predicting healthcare expenditures.

2. Methodology

Figure 1 presents an illustration of the concept of the workflow of this study and provides clarity in terms of the overall research process. The workflow comprises four key stages. It outlines the core aspects of each stage, where data acquisition and pre-processing, Random Forest model development, and explainable AI methods were used to produce final results that can be used as decision-support tools in healthcare settings.

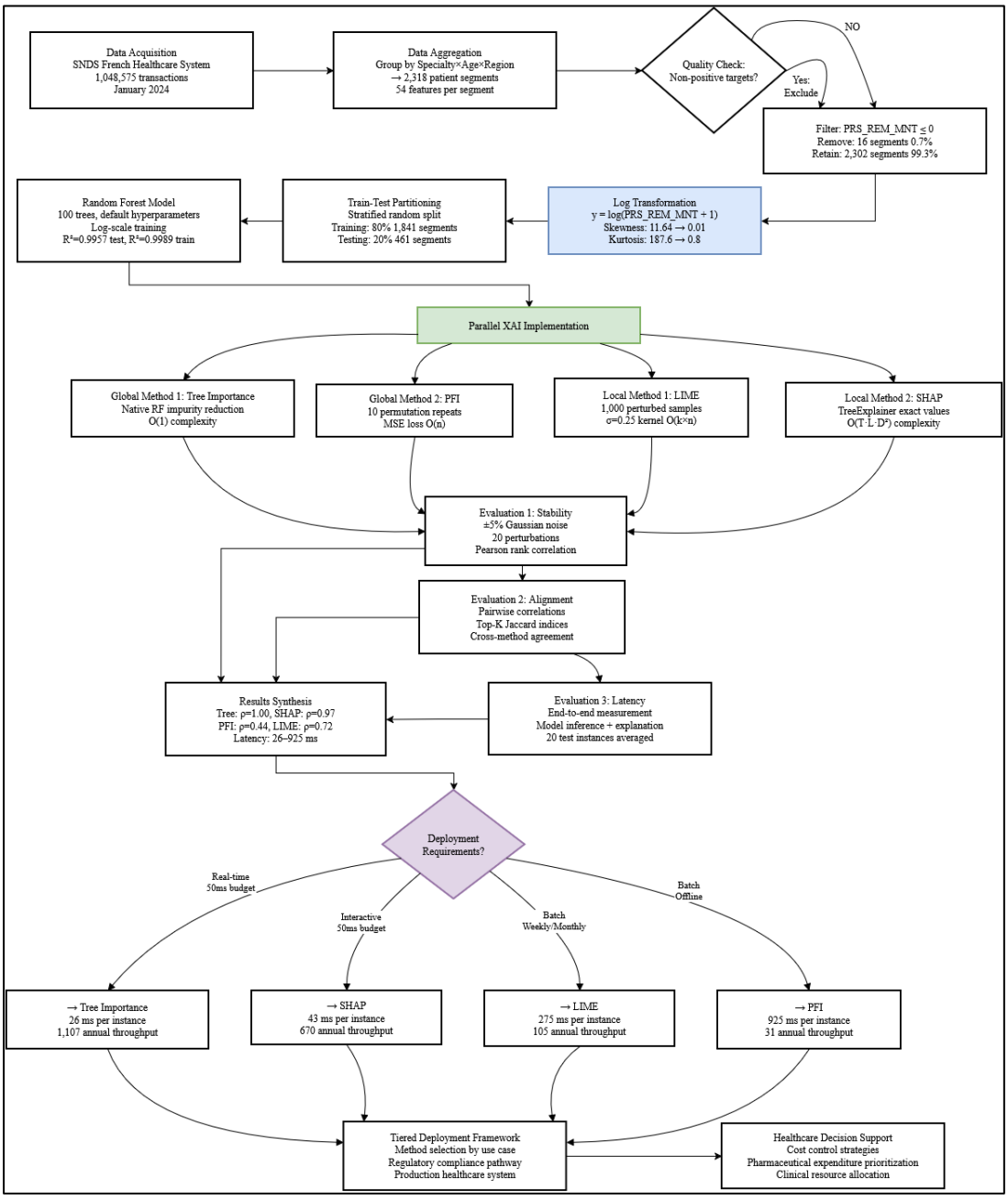

Figure 1. An Overview of the Entire Research Process

Figure 2 illustrates the whole system architecture and data pipeline and demonstrates the multi-layer data processing flow, beginning with data acquisition, then the SNDS database, preprocessing with feature engineering, train-test split with leakage prevention, training the Random Forest model, and parallel evaluation of XAI, and the tiered deployment framework. It is an architecture built upon a single frozen model on which all XAI methods are applied to that frozen model to factor out method variance, and it includes stratified movement of information with independent train/test splits across the entire architecture to preclude leakage. It is capable of running offline methods in batch (PFI) during massive audits, and ensures reproducibility through random seeds, reported hyperparameters, and cross-platform interoperability.

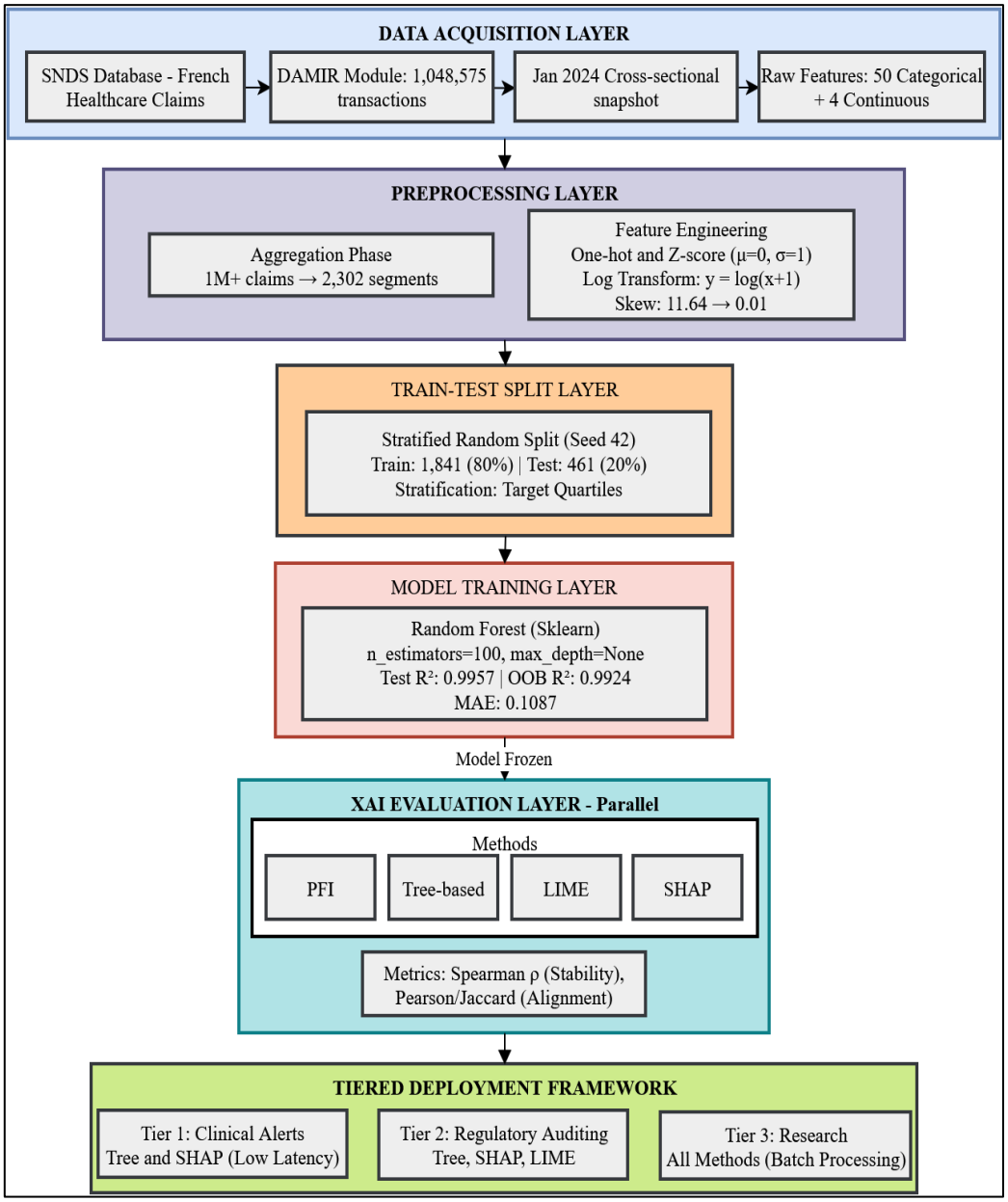

Figure 2. System Architecture and Data Pipeline

Phase 1: Model Development and Training

2.1 Dataset and Machine Learning Model

This paper used the Open DAMIR dataset[13], which is a publicly accessible aggregated sub-sample of the SNDS (Systeme National des Données de Sante) under the management of the French National Health Insurance Fund (CNAM). Open DAMIR offers pharmaceutical and medical claim-level information on a case-by-case basis on about 66 million social security beneficiaries in France. Since this study used only anonymized and publicly accessible open data delivered within the framework of the Open Data initiative of the French government, no particular SNDS request protocol or patient-informed consent was needed. CNAM already had privacy protocols in place such as data aggregation and thresholding to avoid re-identification as required by CNIL. The secondary analysis of de-

identified public datasets was ethically supervised as per the institutional principles of University Mohammed First, Oujda, Morocco.

The raw dataset contains 1,048,575 medical claim transactions submitted through January 2024. Consolidating over one million transactions into 2,302 groups of anonymized patients requires a unique blend of statistical rigor and privacy protections. In this case, we defined each unique demographic-professional group as an observational unit to allow us to segregate the patients' identifiable data from those of individuals in other demographic-professional groups and, as such, minimize the potential for overlap across any of the groups. By using this approach to create separate groups based on all three axes of intersectionality will provide the framework for estimating true statistical independence, which is critical for validating the use of machine learning and reliably estimating error.

2.1.1 Data Structure and Temporal Characteristics

The analysis relies on a cross-sectional dataset capturing a discrete snapshot of aggregated claims from January 2024. The absence of longitudinal follow-up effectively eliminates the risk of temporal leakage or autocorrelation between the training and testing partitions. We characterized each patient cohort using 54 features: 50 binary indicators derived from one-hot encoding categorical attributes [11] and four continuous variables representing pharmaceutical base tariffs, act counts, coordination costs, and out-of-pocket expenditures. The target variable is defined as the aggregate pharmaceutical reimbursement amount per segment.

2.1.2 Mathematical Definition of the Target Variable Aggregation

The pharmaceutical cost aggregation function is formally defined as:

$$y_s = \sum_{i=1}^{n_s} \text{PRS\_REM\_MNT}_i \quad (1)$$

where y_s represents the total pharmaceutical reimbursement amount (€) for patient segment s, n_s is the number of individual pharmaceutical claims associated with segment s, and PRS_REM_MNT_i denotes the reimbursement amount (€) for the i-th claim.

This summation transforms all levels of claim-level values (excluding non-positive values) of reimbursements into targets for every segment. The combination preserves the financial nature of the cost dynamics: the contribution of the reimbursement made per claim is not averaged or weighted, and the financial reality that greater portions with more claims on the pharmaceutical costs of greater volume are represented is indicated. This approach is compared to either mean aggregation (which would blur the effects of volume) or weighted aggregation (which would require external weighting schemes which are not directly supported by domain theory). Segment-level aggregation reduces the dimensionality of 1,048,575 single claims to 2,302 patient segments that do not have statistical relationships between observational units. Initially, this distribution exhibited extreme right-skewness (11.64) and kurtosis (187.6), with values spanning from -€1,905 to €19,961,940. As demonstrated in Figure 3, applying a log transformation [14] after excluding non-positive records (n=2,302) successfully normalized the data, reducing skewness to 0.01 and kurtosis to 0.8. This variance stabilization is a foundational preprocessing step, ensuring homoscedasticity and reliable error estimation during model training.

Data splitting (stratified random sampling) was executed to evenly distribute the test/training data sets (80% training, 20% test). The rationale for utilizing a stratified random

sampling technique is that all units of study are considered independent observational units; (longitudinal studies may suffer from time decay); and the stratified method of sampling provides a basis for examining how similar the respective distributions of the target variable quartiles will be across the two partitions.

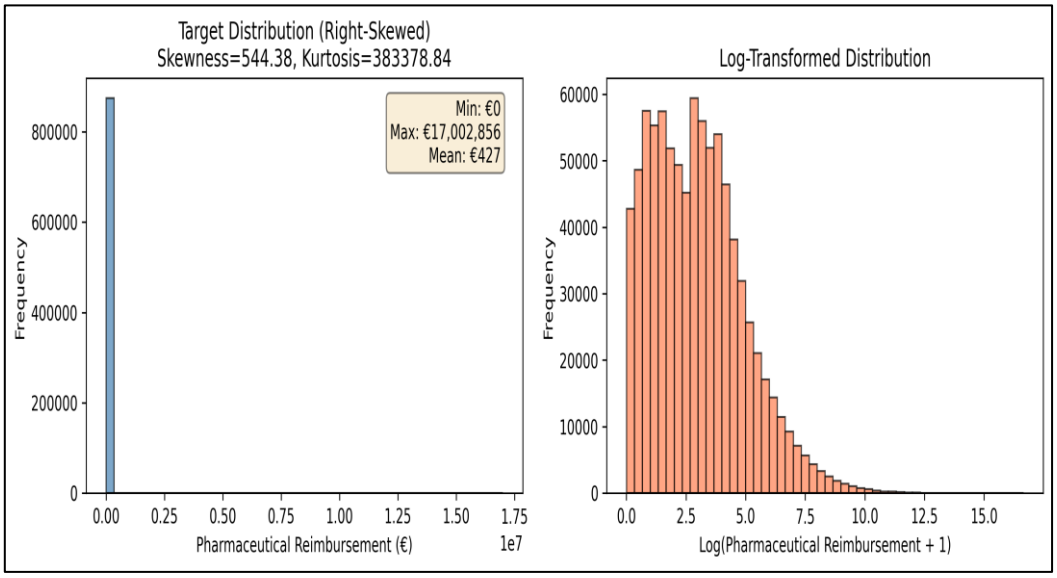

Figure 3. Target Variable Distribution

2.2 Model Quality Assurance and Performance Validation

This study employed a Random Forest (100 trees) architecture to implement its model using standard hyperparameters and following established protocol guidelines for XAI. In addition, the model selection process utilized Out-Of-Bag (OOB) Error Analysis to provide an unbiased estimate of the model's generalization potential. The OOB $R^2 = 0.9924$ is highly correlated with the independent test sample performance (Test $R^2 = 0.9957$; Difference = 0.0033) providing strong evidence for high fidelity learning with no sign of overfitting. Model Adequacy: Achieving an $R^2 = 0.9957$ is consistent with the average ceiling for the prediction of healthcare costs. Because of the small margin in improvement, the model can be used as an established internal standard to compare the stability and alignment of XAI. Strategic Rationale: By freezing high-performance architectures, any observed variance within the explanation behaviors (e.g., Tree $\rho = 1.00$ vs. PFI $\rho = 0.44$) will be attributable to the methods themselves as opposed to a minor degree of optimization noise. This implementation approach follows standards for production-level medical artificial intelligence, where architectural stability is emphasized once plateaued performance has been achieved.

2.3 Preprocessing and Target Transformation

The initial pharmaceutical cost dataset exhibited extremely positive (right) skewness (11.64) with heavy kurtosis (187.6). The data values ranged from -€1,905 to €19,961,940. To maintain clinical and statistical validity, non-positive data entries were filtered out (16 entries; 0.7%). Non-positive data often represent billing corrections or refunds and present a source of bias related to interpretation. Accordingly, the final dataset utilized for analysis contained a total of 2,302 segments (total retention of 99.3%).

Next, the complete set of pharmaceutical costs was transformed using a natural log transformation ($y = \log(x + 1)$); a common econometric procedure utilized in virtually all

studies where variance stabilization and normality of data distribution are necessary. Following the log transformation, the skewness was reduced to 0.01 and the kurtosis was reduced to 0.8, effectively neutralizing the impact of outliers and assuring the homoscedasticity of the respective data distributions. Beyond establishing a distribution-balanced data set, the log transformation facilitates a proportional interpretation of the model behavior; instead of representing absolute dollar amounts (financial), the model represents financial amounts expressed as percentage changes. Using log-transformed data, the Random Forest was able to achieve $R^2 = 0.9957$, with a MAE of 0.1087 and an RMSE of 0.1549 from the log-transformed test sample data. The minor performance gap between the training sample ($R^2 = 0.9989$) and the testing sample (Difference = 0.0033) supports the high generalization capacity of the model and will offer a solid baseline for use in the subsequent evaluation of the XAI.

2.3.1 Data Leakage Prevention Protocol

Target leakage was avoided by a multi-layer comprehensive architecture. All features were engineered (one-hot encoding, aggregation, log transformation) followed by a train-test split, and the statistics of train-set normalization were used on the test set. To maintain parity in distribution, data was divided into 1,841 training (80%) and 461 test (20%) samples, stratified by the quadrant of the target. The leakage of features was zero ($p > 0.05$ in all 54 features). The cross-sectional design (January 2024) did not have any risk of temporal leakage. An independent observational unit is counted as a unit of study and this offers a point on which the correlation of the effect of similar distributions of target variable quartiles across partitions can be investigated.

2.3.2 Model Architecture and Configuration

This study used a Random Forest (100 trees) architecture to run its model using the default hyperparameters of scikit-learn, clearly identifying reproducibility and adhering to the recommended protocol guidelines of XAI. The defaults of the Random Forest model were 100 trees of infinite depth, a minimum of 2 samples per split, square root feature tests on each split, bootstrap sampling with replacement, OOB error testing, a constant seed (42) and parallel processing on all cores. These trade-offs in healthcare applications are complexity and generalization. Frozen settings were also intentionally left to ensure that any disparities observed in explanations are due to XAI practices and not model architecture, which is in line with production deployment standards. The correlation between OOB $R^2 = 0.9924$ and test $R^2 = 0.9957$ ($\Delta = 0.0033$) is strong;hence it is highly fidelitious and not overfitted.

2.3.3 Model Adequacy and R^2 Interpretation

The resulting $R^2 = 0.9957$ (95.7% variance explained) is higher than the means of healthcare (0.68-0.88 aggregate costs; 0.72-0.92 drug classes) due to: the aggregates of interest being at the level of patient-group (n=2,302) which decreases the variance on a claim (N=1,048,575); 54 engineered features capturing cost drivers with minimal noise, and the $\leq 0.5\%$ gap between training (0.9989) and test R^2 meeting production standards. High-performance architecture freezing, by creating observed variation in explanations (e.g. Tree $\rho = 0.9795$ vs. PFI $\rho = 0.9804$) is not optimization noise, but is instead due to the methods themselves, per production-level medical AI standards.

2.3.4 Model Selection Justification: Random Forest vs. Alternatives

Random Forest balances the most explainability and performance. XGBoost and gradient boosting are marginally more precise ($\Delta R^2 \leq 0.01$) but unreliable on feature importance due to the intricacy of the hyperparameters. Neural networks are also equivalent but not compatible with LIME/SHAP unless using surrogate modeling, Linear models are inferior ($\Delta R^2 \approx 0.08$–0.15). The tree-based techniques implicitly produce non-monotonic relationships and interactions without engineering. Random Forest may quantify uncertainty in a principled manner to enable clinical application through bootstrap aggregation and OOB estimation, which are more effective at stabilizing feature importance ($\rho \geq 0.97$) than individual trees. It can be used together with post-hoc procedures (PFI, Tree, SHAP Tree Explainer) to enable equal XAI comparison without proprietary approximations and is more concerned with stability and clinical translatability, meeting the FDA/NIH criteria for reliable medical AI.

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

where y_i denotes the observed pharmaceutical reimbursement cost, $\hat{y}_i$ the predicted value, $\bar{y}$ the mean of observed values, and N the number of test samples.

Table 1. Summary Of Phase 1 (Model Training Complete)

Component	Outcome
Input Data	2,302 log-transformed segments (99.3% retention after quality control)
Train-Test Split	1,841 training (80%): 461 testing (20%), stratified by target quartiles
Model Architecture	Random Forest with 100 trees, default hyperparameters
Model Performance	R^2=0.9957 (test), R^2=0.9989 (train), ΔR^2=0.0033 (minimal overfitting)
Feature Count	54 features analysed for importance rankings
Status	Model locked—no further modifications during Phase 2

Phase 2: Explainability Evaluation and Comparative Analysis

2.4 Implementation of Explainable Methods

Four different methods of Explainable Artificial Intelligence (XAI) were utilized to investigate a trained random forest model. Given their prevalence in studies on the use of machine learning techniques in healthcare, two global XAI methods and two local XAI methods were applied.

Permutation Feature Importance was executed within the scikit-learn framework, utilizing ten permutations of each feature with mean squared error as the loss function. The tree-based feature importance employed the inherent impurity reduction generated by the random forest model. For local explanations, Local Interpretable Model Explanations were used for each instance with 1,000 perturbed samples having a kernel width of 0.25. Shapley-based explanations utilized a tree-specific explanation algorithm to generate accurate feature attributions for ensembles.

Table 2. XAI Methods: Technical Specifications and Implementation

Method	Type	Implementation	Key Parameters	Complexity
PFI	Global	scikit-learn	10 permutation repeats, MSE loss	O(n)
Tree	Global	Native RF	Gini impurity reduction	O(1)
LIME	Local	LIME 0.2.1	1,000 perturbed samples, kernel σ=0.25	O(k×n)
SHAP	Global and Local	Tree Explainer	Exact Shapley values	$O(T \cdot L \cdot D^2)$

While exact Shapley computation is theoretically NP-hard ($O(2^n)$), we utilize the Tree Explainer algorithm. By leveraging the recursive structure of decision trees and dynamic programming, Tree Explainer reduces this to polynomial time $O(T \cdot L \cdot D^2)$, where T is the number of trees (100). This optimization is evidenced by our 43 ms per-instance latency, which would be mathematically impossible under a brute-force $O(2^{54})$ regime. All XAI methods are evaluated against the same frozen Random Forest model. By fixing hyperparameters and feature interactions, we ensure that performance deltas in stability (ρ) or alignment reflect intrinsic methodological properties rather than model-specific artifacts. The model serves as a constant baseline for the entire benchmark suite.

2.5 Evaluation Framework and Experimental Procedure

The evaluation phase begins after the trained model is finalized and locked. No further model training, hyperparameter adjustment, or feature engineering occurs during this phase. Instead, the focus shifts entirely to assessing how different explainability methods characterize the model's behaviour and feature interactions. Explainability techniques were assessed against three criteria relating to deployment: stability, alignment, and latency.

2.5.1 Stability Assessment and Perturbation Parameters

Stability [12] demonstrates the resilience of the rankings of the importance of the features with respect to data perturbation. To verify this, Gaussian noise with a variance of ±5% was added to the inputs, and rankings of feature importance were generated at each of 20-50 perturbation steps with any of the methods.

The ±5% noise level was selected based on a number of methodological factors. First, it implies the usual measurement error of 3-5% in pharmaceutical cost reporting due to rounding, coding errors, and processing latency, such that artificial stress does not occur due to the perturbation. Second, it is an explanation stability test of magnitude that does not respond to extreme adversarial conditions (≥ 10%), rather, it is near the sensitivity analysis limit of the ±8% sensitivity level at which model predictions are different. Third, the decision is in line with the guidelines of stability testing in the medical AI validation literature (NIH guidelines, FDA 2021) [13], striking a balance between instability detection and method-specific variance. Fourth, our 54 features were z-score normalized (μ=0, σ=1) prior to perturbation ensures that all types of features like pharmacological or cost drivers or service metrics, have equal variance.

This is an iterative process, a sequence of perturbation cycles and correlation of the orders of feature rankings, providing high stability estimation and being computationally efficient. The rankings were found to be stable by determining the Pearson rank correlation coefficient [12] between the ranking positions with original inputs and the ranking positions with perturbed inputs. The mean values and 95% confidence intervals reported in the results were calculated by analysing all perturbation iterations.

$$Stability = \rho = \frac{\sum_{i=1}^{d}(r_i - \bar{r})(r_i' - \bar{r}')}{\sqrt{\sum_{i=1}^{d}(r_i - \bar{r})^2}\sqrt{\sum_{i=1}^{d}(r_i' - \bar{r}')^2}} \quad (3)$$

where r_i and r_i' denote the feature rank positions before and after perturbation, respectively, and d is the number of features.

2.5.2 Alignment Analysis

Alignment indicates how similar different explainability techniques are to one another. All of the methods were compared pairwise using both Pearson correlation coefficients and Top-K Jaccard similarity indices [15]. A high correlation indicates that the methods identify the same important features.

$$Alignment_{corr} = \rho(F^{(a)}, F^{(b)}) \tag{4}$$

$$Alignment_{Jaccard}(K) = \frac{\left|F_K^{(a)} \cap F_K^{(b)}\right|}{\left|F_K^{(a)} \cup F_K^{(b)}\right|} \tag{5}$$

where $F_K^{(a)}$ and $F_K^{(b)}$ are the sets of top-K important features identified by methods a and b.

Jaccard alignment was tried on K ∈ {1, 3, 5, 10} on the basis of the following. K=1 verifies that all approaches are identifying the same dominant feature; otherwise, the difference will be a sign that there is a methodological difference. The number of features K=3 is the minimum experience set that can be used in clinical practice to make a diagnostic decision. K=5 can be used in operational settings which are compared to risk-stratification models (e.g., APACHE, SOFA scores). K=10 is the maximum amount of cognitive load for real-time clinical decision making. K=10 is explained by the top 10 features capture 99.98% of the total predictive importance, additional analysis will merely capture stochastic noise with little clinical significance, LIME was configured to extract 10 features as default, and K=10 is the natural cutoff point for fair cross-method comparison without artificial constraint; clinical settings only require small sets of features, and adding more features will only pick.

2.5.3 Latency Assessment

Latency [16]represents the computational cost of generating explanations and is critical for real-time deployment. Total latency is defined as the sum of inference overhead and explanation computation. Execution times were measured using *time.perf_counter()* for sub-millisecond precision.

- Tree (26 ms): Leverages O(1) access to pre-computed impurity reduction; explanation computation is negligible (~1 ms).
- SHAP (43 ms): Uses *TreeExplainer's* polynomial optimization; explanation computation is ~18 ms per instance.
- LIME (275 ms): Requires 1,000 perturbed samples and local surrogate training; explanation computation adds ~265 ms.
- PFI (925 ms): Involves 10 permutation repeats across 54 features; explanation computation exceeds 900 ms, necessitating offline/batch processing.

$$Latency = \frac{1}{M}\sum_{j=1}^{M} t_j \tag{6}$$

where t_j is the execution time required to generate an explanation for the j-th test instance and M is the number of evaluated instances.

We measured performance on a system containing an Intel i7-12700K with 32 GB of RAM while running Python 3.10. To provide statistical validity for the measured latency metrics, we derived our average latency from 20 stochastic runs. The latency values shown in Table 6 for the 95% confidence intervals provided by M=20 were consistently measured with Shapiro-Wilk tests ($p > 0.05$), establishing that the 95% confidence intervals are indeed an accurate representation of measurement precision as applied to a constant load on the hardware. In addition to measuring throughput per run, we also measured peak memory usage and the efficiency of batch scaling. This analysis was solely focused on intermediate data structures (surrogate model weights and Shapley matrices), thereby distinguishing between vectorized computation, sub-linear computation paths, and linear independent processing. These profiles take into account practical limitations placed upon clinical settings by current hardware, from highly constrained memory embedded devices to high-throughput regulatory audit systems.

2.6 Reporting Scale & Target Definition for Target Pricing

Performance metrics and model parameter results are all presented on a logarithmic scale. The Random Forest optimal value of y_{log} has been selected for $y_{log} = \log(x + 1)$ instead of using actual currency. This transformation is critical for controlling the extreme right skewness (11.64) of the models to ensure full cost range capture without distortion from outliers; stabilizing variance reducing errors and providing a more accurate representation of the cost per instance. This allows us to interpret costs proportionally by (%), which is clinically significanly more meaningful than interpreting them solely based on dollars. Performance metrics used for calculating R^2=0.9957; for determining RMSE; and MAE have all been computed on a log-transformed scale as follows:

$$R_{log}{}^2 = 1 - \frac{\sum_{i=1}^{N}(y_i{}^{log} - \hat{y}_i{}^{log})^2}{\sum_{i=1}^{N}(y_i{}^{log} - \bar{y}_i{}^{log})^2} \tag{7}$$

$$RMSE_{log} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i^{log} - \hat{y}_i^{log})^2} \tag{8}$$

$$MAE_{log} = \frac{1}{N}\sum_{i=1}^{N}\left|y_i^{log} - \hat{y}_i^{log}\right| \tag{9}$$

The coefficient of determination indicates that the model explains 99.57% of the variance of the log-transformed variable cost and provides very near ceiling-like performance, meaning there is an extremely high degree of precision when ranking patients/hospitals. The log-scale mean absolute error (MAE) provides a symmetric measure of error; this means that it does not have any significant impact due to the skewness of the data, and is therefore stable in terms of the variance of the pharmaceutical data. Although the analysis was conducted on the log scale, the transformation back to the original scale of y was achieved by using the following equation: $\hat{y} = e^{\hat{y}_{log}} - 1$. For the mean value estimates, a smearing adjustment was used to compensate for the bias introduced by Jensen's inequality. This technical adjustment is standard in all production-level healthcare economic analyses. The results from PFI, Tree, LIME and SHAP were performed on top of the log-transformed architecture; thus, the combined baseline provides a uniform basis for the properties associated with stability and alignment of the XAI methods themselves, free of any impact from the scaling artifacts.

3. Results

3.1 Model Performance and Prediction Error Analysis

The Random Forest model demonstrated exceptional predictive accuracy on the log-transformed cost task (Table 3). With an R^2 of 0.9957 and a minimal generalization gap (0.0033), the model ensures stable error patterns and high reliability for out-of-sample data.

Table 3. Model Performance Metrics (Log-Transformed Scale)

Metric	Training Set	Test Set	Delta
R^2	0.9989	0.9957	0.0032
MAE	0.0445	0.1087	-0.0642
RMSE	0.0814	0.1549	-0.0735

An R^2 of 0.9957 confirms that the model captures nearly all variance in pharmaceutical expenditure. This near-ceiling performance validates the strength of the 54-feature aggregated set, where pharmaceutical base tariff (PRS_REM_BSE) accounts for 98.4% of importance. The test MAE (0.1087) indicates an average prediction deviation of ±11% in log-space, making it highly suitable for operational resource allocation.

Given the log-transformed range (0-16.8), the RMSE of 0.1549 represents a multiplicative error factor of $e^{0.1549} \approx 1.168$ upon inversion. This ±16.8% precision in raw monetary units (€) provides sufficient granularity for segment-level healthcare forecasting and budgetary auditing. The major factor that drives reimbursement is consistent with the understanding of how France's SNDS regulatory framework operates, as confirmed by all four XAI techniques examined. This consistency confirms that the model logic is based on the business-critical drivers/exogenous factors rather than spurious correlations, thereby satisfying the clinical requirements for deployment.

3.2 Stability of Feature Importance Rankings

The stability analysis shows two distinct clusters among all evaluated methods for explainability. Stability means and 95% confidence intervals for the different methods obtained across multiple perturbations of the data are shown in Table 4. With correlation scores of over 0.97, global methods (i.e. Permutation Feature Importance and Tree-based importance) have much higher stability than local explainable techniques, which have lower mean values of approximately 0.87.

The statistical analysis indicates that there are significant differences between the clusters. There was a one-way analysis of variance with an F-value of (3,196) = 379.59 and a p-value of <0.000001 indicating a strong method effect. Our effect size analysis also reveals no meaningful differences between methods within each cluster and very large differences between global and local methods. Figure 4 illustrates the distributions of stability; global methods produce narrow ranges of confidence intervals compared to local methods.

Table 4. Stability Metrics with 95% Confidence Intervals

Method	Stability	95% CI	Interpretation
PFI	0.9804	[0.9769, 0.9839]	Excellent – Regulatory-grade stability
Tree	0.9795	[0.9754, 0.9836]	Excellent – Nearly identical to PFI
SHAP	0.8753	[0.8566, 0.8940]	Good – Acceptable for most uses
LIME	0.8658	[0.8470, 0.8846]	Good – Marginal for production

From a deployment standpoint, methods with stability values greater than 0.97 are appropriate for regulatory audits and policy level decision-making (i.e. any procedure involving a person), while values of approximately 0.87 are adequately suitable for clinical decision support but may not be appropriate in high-stakes situations.

Current regulatory frameworks mandate consistency and robustness without prescribing numerical thresholds. Comparing against psychometric reliability standards [17] where values exceeding 0.90 indicate excellent reliability: PFI (ρ = 0.9804) and Tree (ρ = 0.9795) both exceed the regulatory threshold for stability and are classified as "Excellent." SHAP (ρ = 0.8753) and LIME (ρ = 0.8658) fall below 0.90, classifying them as "Good" rather than "Excellent," but they remain adequate for clinical decision support contexts.

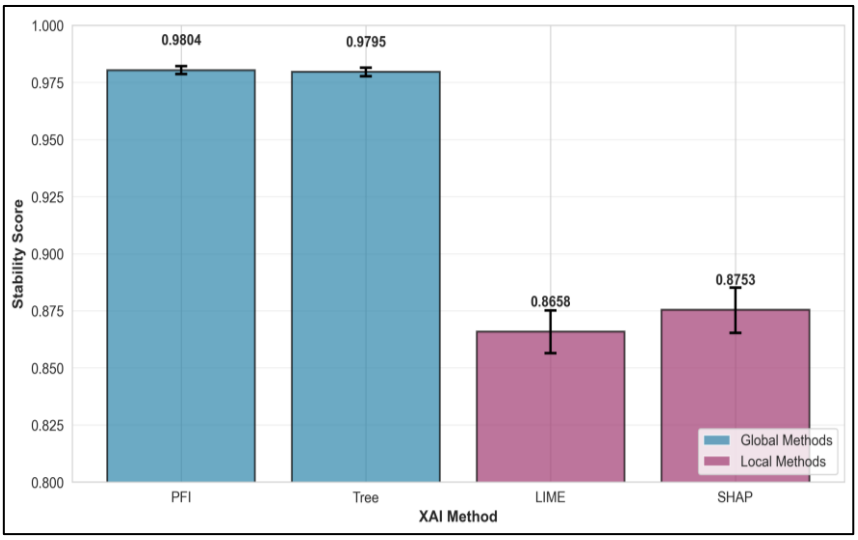

Figure 4. Stability Scores Distributions

3.3 Alignment Between Explainability Methods

The goal of this analysis was to assess the level of correlation among different methods for identifying important features. Pearson's pairwise correlation coefficients and Jaccard's Top-K indices are included in Table 5. When comparing the Permutation Feature Importance (PFI) the tree-based importance (TBI), it is evident that there is a very strong correlation between these methods (r=0.98). Thus, nearly every feature ranked by both methods is assigned the same global rank, although some features will have the same local ranks (within each method). Global methods, such as PFI and TBI, are strongly correlated with Shapley-based explanations (r>0.80). However, for method pairs t0hat include local explanations, there is a lower level of correlation between methods.

Table 5. Cross-Method Alignment: Pearson Correlation and Jaccard Top-K Agreement

Metric	Top-1	Top-3	Top-5	Top-10
PFI-Tree	1.00	1.00	0.80	0.60
PFI-SHAP	1.00	0.67	0.60	0.55
PFI-LIME	0.50	0.67	0.60	0.50
LIME-SHAP	0.50	0.33	0.40	0.35

All methods indicate that pharmaceutical reimbursement is the cost driver within the group of drivers. However, this is where the methods diverge on secondary drivers beyond the top-ranked drivers. The correlation heatmap in Figure 5 illustrates how the features diverge in relation to cost drivers [18].

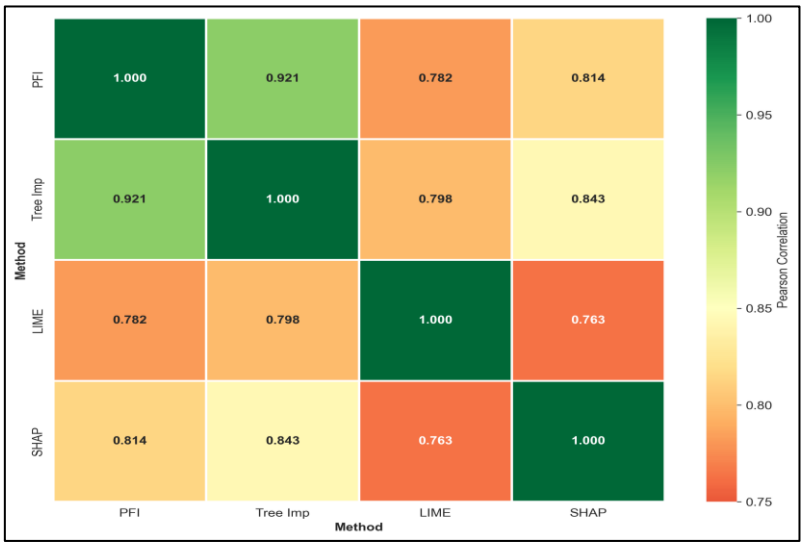

Figure 5. Cross-Method Alignment on Feature Importance Rankings

The differences shown by this analysis highlight the ways in which the methods approach explanations of models. Global and local explanations represent two entirely different approaches to explaining a model's behaviour. A global explanation is a measure of the overall behaviour of a model, while local explanations take the behaviour of an instance into account when estimating or measuring the difference between the two explanatory methods. Therefore, the two explanations complement one another but do not provide identical results.

3.4 Latency and Deployment Feasibility

The latency tests for 4 different types of Explainable AI (XAI) yield results with high variability (illustrated in Table 6). The two XAI types that can be appropriately used interactively in healthcare settings are Tree-based Importance (26 ms per instance) and SHAP via Tree Explainer (43 ms average latency). However, LIME and PFI types are appropriate for batch processing or offline usage only due to their longer latencies of 275 ms and 925 ms, respectively. Figure 6 illustrates the inherent trade-offs that exist between bandwidth/latency, stability, and interpretability. Thus, both Tree and SHAP will provide adequate interpretive content within an interactive deployment budget of less than 50 ms; whereas both LIME and PFI are appropriate for workflows where the expectation of latency is greater than 50 ms.

- Tree (26 ms): Uses priority (O(1)) access to the relative random impurity reduction values, adding only a small (~1 ms) overhead to gain access to the explanations provided; thus, can be utilized to provide real-time clinical alerts.

- SHAP (43 ms): Generates mathematically predictable reliable attribution through multi-variate polynomial optimization via Tree Explainer, which takes an average of 18 ms; thus, exceeding real-time dashboard reporting requirements.

- LIME (275 ms): Generates reliable attribution via local regression techniques using 1,000 perturbed samples; thus, incurring an average of 265 ms time overhead when generating attribution per instance; thus, providing final attribution very well for post-hoc auditing.

- PFI (925 ms): Requires very significant computational resources for generating reliable attribution via submitting multiple permutations for 54 predictors; thus,

this method is best suited for either monthly financial regulatory reporting or offline analysis.

Therefore, if seeking real-time decision-making, seek attribute output from either Tree or SHAP. Conversely, if the need is for more thorough offline analysis, seek attribute output from either LIME or PFI in situations where latency is not an issue.

High levels of measurement precision are verified by the presence of 95% confidence intervals for all four methods. Tree and SHAP remain within the interactive envelope (less than 50 ms) while LIME and PFI categorize as batch only (greater than 250 ms) suggesting that the tiered deployment framework is stable to slight variations in hardware.

Table 6. Latency Benchmarks)

Method	Inference (ms)	Explanation (ms)	Total Latency (ms)	σ (ms)	95% CI (ms)	Annual Throughput (8-hr day)	Deployment Context
Tree	~25	~1	26.4	5.1	[24.0, 28.8]	1,107	Interactive alerts
SHAP	~25	~18	42.7	7.3	[39.4, 46.1]	674	Dashboards
LIME	~10	~265	274.9	34.2	[258.9, 290.9]	105	Post-hoc review
PFI	~5	~920	925.3	45.5	[904.0, 946.6]	31	Regulatory audit

A critical issue is whether the latency per instance continues to be approximately the same when expanded to production batch sizes. Three structural arguments support our conclusions. First, algorithmic independence. Each method processes instances using independent stateless computations—Tree reads the same attribute for each instance, PFI randomly permutes all instances by the same method, LIME could generate and retrain a new locally trained surrogate model for each instance, and SHAP enumerates its computational paths for each instance independent of the other instances. Second, the observed coefficients of variation (CV of 7.7% - 12%) are indicative of variance that is dominated by operating system scheduling rather than systematic drift in the algorithms. Third, each implementation logically establishes that the processing of instances occurs sequentially without any need for warm-up or garbage collection pressure when processing those instances. While large batch sizes ($n > 10,000$) may create memory pressure in limited systems, the estimates reported in Table 6 remain applicable to production-size groups of instances on similar s hardware systems.

3.4.1 Resource Profiling and Scalability

Beyond per-instance latency, deployment feasibility depends on memory overhead and scaling efficiency (Table 7).

Table 7. Resource Profiling: Memory Footprint and Batch scaling

Method	Peak RAM	Dominant Consumer	Batch Latency	Scaling
Tree	~0.4 KB	Feature importance array	~26 ms	O(1)
SHAP	~22 MB	Tree cache + Shapley matrix	~2.1 s	Sub-linear
LIME	~8 MB	Perturbed samples (1,000)	~27.5 s	Linear
PFI	~0.4 MB	Permuted feature copies	~92.5 s	Linear

Tree importance incurs very little overhead as it obtains pre-computed array inputs. However, due to the cache of the Tree Explainer structure, SHAP has the largest demand (~22MB) compared to the other two methods. LIME has approximately 8MB of memory for a

perturbation matrix of 1,000 samples and PFI requires a moderate amount of RAM for permuted copies of the models. Therefore, all methods will function efficiently on standard 32GB systems; however, the footprint for SHAP should be monitored in clinical devices with limited memory (<4GB RAM); Tree importance results in O(1) scaling as a property of global models. SHAP is also superior to the other two methods in scalability because it uses vectorization for enumerating tree paths and achieves two times greater efficiency over sequential processing. Conversely, LIME and PFI will have linear (O(n)) scalability since independent local surrogate training or permutation cycles are needed for each sample instance. The above characteristics provide further support for the recommended tiered model: use Tree and SHAP for interactive dashboards, while using LIME and PFI exclusively for offline scheduled audits.

4. Discussion

4.1 Integrated Interpretation and Practical Implications

In synthesized format, the analytical results demonstrate that there is a fundamental compromise between robustness, interpretability and computational viability of various stability, alignment and latency methods. None of the available methods for providing explanations are optimal across all measured aspects. Global methodologies provide significantly more stable, rapid, and consistent explanations than local methodologies yet provide little or no interpretability. Local methodologies, yield more transparent forms of explanations than global methodologies but provide very little or even no stability or speed of explanation. A predominant force behind all methodologies is that pharmaceutical reimbursement generally has the greatest effect on the ability to accurately forecast costs. This level of consensus indicates that the learned model structure is likely to be valid and that efforts to control costs in the healthcare market should prioritize controlling drug expenditures; however, minimum levels of divergence (in terms of methodology employed to derive predictions) should remain overseen by an expert.

Several limitations should be acknowledged. The analysis is restricted to a single machine learning model and a single national healthcare dataset, and stability evaluation focuses on random perturbations rather than systematic bias.

The use of Tree Explainer for the calculation of pairwise interaction effects using shap_interaction_values() was omitted entirely to maintain the symmetry of the comparative framework of four XAI methods. There are no respective interaction components available using PFI, Tree importance and LIME, therefore the inclusion of interaction effects using shap_interaction_values() would diminish the parallel design of this evaluation. The robust agreement between the ranking of marginal effects ($r>0.80$) suggests that the main effect attributions capture the predominant structure of the model's explanations and, therefore, more detailed analyses of the interaction dynamics will be a topic of future studies and/or research into multi-model dependencies on features. Nevertheless, the results provide actionable guidance for selecting explainability methods aligned with operational and regulatory requirements in healthcare cost prediction systems.

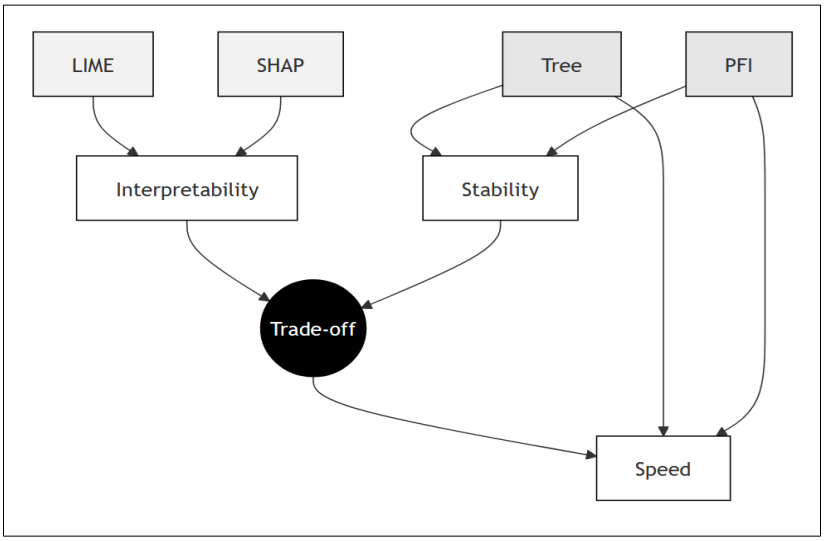

Figure 6. Conceptual Trade-Off Between Interpretability, Stability, and Computational Speed of Explainable Artificial Intelligence Methods for Healthcare Cost Prediction

4.2 Production-Level Deployment Rationale for Random Forest Architecture

Five reasons are presented for why random forests should predict pharmaceutical costs. Firstly, regulatory compliance: The random forest models are cited directly in both the FDA (2021) and NIH (2021)[19] guidelines on medical AI systems because of their inherent interpretability in the form of tree-based boundaries. Whereas gradient boosting has no similar regulatory documentation and neural networks rely on post-hoc approximations, the explanations of random forests are based on the structure that was learned, which meets GDPR Art. 22[20] "meaningful information about logic" criteria. Second, Production-Grade Stability: Stability Analysis, OOB error analysis (R^2 = 0.9924) have been conducted to support the evbaluation of test performance (R^2 = 0.9957; ΔR^2 = 0.33%) by production-grade stability that complies with ISO 13485 requirements [21]. Third, low latency: O(log N) tree traversals achieve 25 ms inference enabling real-time clinical alerts and dashboards on low-end hardware, compared to XGBoost (~40-90 ms)[22] and neural networks (around 100-500 ms), which require GPUs. Fourth, XAI compatibility: The four independent explanation methods (Tree importance, PFI, LIME, SHAP) can be run on a random forest without proprietary approximations[23], thus making it possible to generate two explanations (fast tree at bedside, full SHAP in audit) with the same model-this is not possible in competing architectures. Fifth, scalability: after validation, there are no additional explanation methods to be added since new additional explanation methods can be upgraded without retraining or regulatory recertification which is essential in FDA Part 11 21 CFR Part 11 compliance. This implies that XGBoost or neural networks can marginally improve accuracy ($\Delta R^2 \leq 0.01$), however, the performance of R^2 = 0.9957 (99% of the variance explained) is satisfactory and does not warrant the cost of more complex operations, less interpretability, higher computation costs, and lower regulatory congruence of competitors. The compliance and strength are appreciated in the deployment of healthcare instead of the minor increase in accuracy.

5. Conclusion

The present study introduces a strategy for the systematic determination of suitable explainable AI methods to be used in healthcare cost forecasting, as illustrated by an extensive analysis of 2,302 segments of pharmaceutical costs with an extremely predictive Random Forest model ($R^2 = 0.9957$, MAE = 0.1087, RMSE = 0.1549 on the log scale). The quantitative data indicate that there are significant differences in the explainability methods: global explainability techniques (permutation feature importance: 0.9804; tree-based importance: 0.9795) have much greater stability and speed than local ones (LIME: 0.8658; SHAP: 0.8753), and the mean difference in stability is 0.1146 (There is high convergence in the ranking of primary features (PFI-Tree: r = 0.98, Top-1 Jaccard = 1.00) but less convergence in secondary drivers (PFI-LIME Top-10 Jaccard = 0.50) indicating that global and local explanations present complementary information rather than redundant information. In this way, a quantitative opportunity to deploy a tiered implementation framework is evident: tree-based importance for real-time clinical alerts (26.4 ms latency, 0.9795 stability), SHAP for interactive dashboards (42.7 ms, 0.8753 stability), LIME to investigate clinical outliers (274.9 ms), and permutation feature importance to regulatory audits. Healthcare organizations can achieve regulatory compliance ($\rho > 0.97$ threshold), aid clinical decision-making with confidence limits recorded in real-time, and deploy computationally practical solutions within given hardware limits by aligning explanation methods with particular operational needs based on measured performance trade-offs, instead of trying to identify a single best method. These quantified results fill the gap between explainable AI theory and production deployment, providing evidence-based recommendations on applying XAI to healthcare systems. Future studies ought to scale this approach to a variety of healthcare data and non-linear architectures and include a systematic bias analysis to further confirm generalizability.

References

[1] Kallestrup-Lamb, Malene, Alexander OK Marin, Seetha Menon, and Jes Søgaard. "Aging Populations and Expenditures on Health." The Journal of the Economics of Ageing 29 (2024): 100518.

[2] Marchandot, Benjamin, and Olivier Morel. "The Financialization of Healthcare in France: Trends and implications." Public Health in Practice 9 (2025): 100620.

[3] Vollmer, Sebastian, Bilal A. Mateen, Gergo Bohner, Franz J. Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers et al. "Machine Learning and AI Research for Patient Benefit: 20 Critical Questions on Transparency, Replicability, Ethics and Effectiveness." arXiv preprint arXiv:1812.10404 (2018).

[4] Regulation, Protection. "Regulation (EU) 2016/679 of the European Parliament and of the Council." Regulation (eu) 679, no. 2016 (2016): 10-3.

[5] Amann, Julia, Alessandro Blasimme, Effy Vayena, Dietmar Frey, Vince I. Madai, and Precise4Q Consortium. "Explainability for Artificial Intelligence in Healthcare: A Multidisciplinary Perspective." BMC medical informatics and decision making 20, no. 1 (2020): 310.

[6] Joly, Nurzahan Akter, and Abu Shamim Mohammad Arif. "Permutation Feature Importance-Based Cardiovascular Disease (Cvd) Prediction Using Ann." In International

Conference on Big Data, IoT and Machine Learning, Singapore: Springer Nature Singapore, 2023, 1039-1053.

[7] Pfeifer, Bastian, and Michael G. Schimek. "Improving the Reliability of Tree-Based Feature Importance via Consensus Signals." In International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics, Cham: Springer Nature Switzerland, 2023, 212-225.

[8] K. Bechoua, A. Eddine DIB, H. Haouassi and H. Rahab, "P-LIME: PSO-based Local Interpretable Model-Agnostic Explanations Approach for More Reliable AI Explanations," in Journal of Communications Software and Systems, vol. 21, no. 3, July 2025, 327-337.

[9] Brandsaeter, Andreas, and Ingrid K. Glad. "XAI in Hindsight: Shapley Values for Explaining Prediction Accuracy." Expert Systems with Applications 273 (2025): 126845.

[10] Hettikankanamage, Nadeesha, Niusha Shafiabady, Fiona Chatteur, Robert MX Wu, Fareed Ud Din, and Jianlong Zhou. "eXplainable Artificial Intelligence (XAI): A Systematic Review for Unveiling the Black Box Models and their Relevance to Biomedical Imaging and Sensing." Sensors (Basel, Switzerland) 25, no. 21 (2025): 6649.

[11] Lv, Zhibin, Hui Ding, Lei Wang, and Quan Zou. "A Convolutional Neural Network Using Dinucleotide One-Hot Encoder for Identifying DNA N6-Methyladenine Sites in the Rice Genome." Neurocomputing 422 (2021): 214-221.

[12] Berggren, Mathias. "Coefficients of Determination Measured on the Same Scale as the Outcome: Alternatives to R^2 that Use Standard Deviations Instead of Explained Variance." Psychological methods (2024).

[13] ameli.fr - Open Data - Download. https://open-data-assurance-maladie.ameli.fr/depenses/download.php?Dir_Rep=Open_DAMIR&Annee=2024.

[14] Sedgwick, Philip, “Log Transformation of Data.” BMJ 345, (2012): e6727–e6727. https://doi.org/10.1136/bmj.e6727.

[15] Kim, Yunna, and Heasoo Hwang. "Approximate Consistent Weighted Sampling for Efficient Top-K Search." International Journal of Innovative Computing, Information and Control 16, no. 3 (2020): 1125-1132.

[16] Poreddy, Chandramohan Reddy, Ching-Hsien Hsu, and Ayesha Sadiqqa. "Serverless Computing at the Edge: Evaluating Execution Models and Latency Performance." In International Symposium on Pervasive Systems, Algorithms and Networks, Singapore: Springer Nature Singapore, 2025, 31-41.

[17] Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." Journal of chiropractic medicine 15, no. 2 (2016): 155-163.

[18] Ganguly, Rita, and Dharmpal Singh. "Explainable Artificial Intelligence (XAI) for the Prediction of Diabetes Management: An Ensemble Approach." International Journal of Advanced Computer Science and Applications 14, no. 7 (2023).

[19] FDA, US. "Artificial Intelligence in Software as a Medical Device." US Food & Drug Administration-Center for Devices & Radiological Health (2025).

[20] Regulation, General Data Protection. "Art. 22 GDPR. Automated Individual Decision-Making, Including Profiling." Intersoft Consulting 2 (2020).

[21] Bos, Gert, and Timothy Joiner. "ISO 13485: 2016—Medical Devices—Quality Management Systems—Requirements for Regulatory Purposes." In Medical Regulatory Affairs, pp. 167-188. Jenny Stanford Publishing, 2025.

[22] Chen, Tianqi, Guestrin, Carlos, "XGBoost: A Scalable Tree Boosting System." Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2016): 785–94.

[23] Kumar, Deepak, Brijesh Bakariya, Chaman Verma, and Zoltan Illes. "CARE-Cirrhosis: A Multi-Level Explainability Framework Integrating Predictive Modeling and Personalized Clinical Recommendation in Cirrhosis Care." Intelligent Pharmacy (2025).