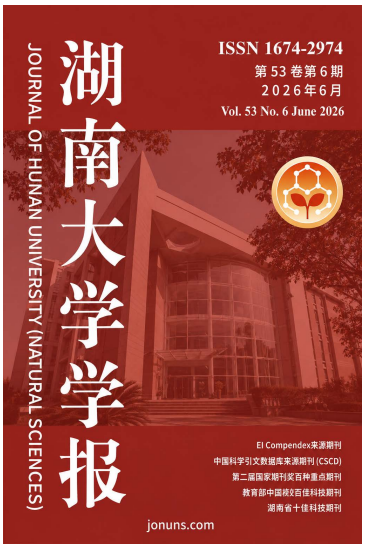

Journal of Hunan University (Natural Sciences)

Vol. 53 No. 6

June 2026

Available online at

https://jonuns.com

Open Access Article

https://doi.org/10.55463/issn.1674-2974.53.6.1

Data-Driven Pharmaceutical Pricing in Developing Markets: A Machine Learning Approach

Abdul Ahad Hassan Farroqi [1], Fahad Hassan Farooqi [2*], Zahoor Ahmad [3], Shah Ameer [3], Waqar Hassan Shahab [4]

[1] School of Economics and Trade, Hunan University, Changsha, China,
[2] Department of Statistics, Government Graduate College, Jhang, Pakistan,
[3] Department of Statistics, University of Sargodha, Sargodha, Pakistan,
[4] Department of Statistics, Quaid-i-Azam University, Islamabad, Pakistan,

* Corresponding author: **fahadg507@gmail.com**

Article History:

Received: April 8, 2026
Revised: May 21, 2026
Accepted: June 5, 2026
Published: June 30, 2026

Abstract: Pharmaceutical pricing in developing markets is often affected by limited transparency, affordability constraints, uneven product availability, and inefficient pricing structures. These challenges complicate evidence-based pricing decisions for pharmaceutical companies, healthcare providers, and policymakers. This study develops a machine learning framework for predicting pharmaceutical prices in the Pakistani market using product-level variables, including company name, pack size, pre-discount price, discount percentage, and product availability. The dataset was obtained from a publicly available Kaggle dataset and included 1,630 pharmaceutical product records. Before model development, exploratory data analysis and preprocessing were conducted to examine price distribution, high-value observations, pack-size variability, missing values, and categorical features. The final post-discount price was used as the target variable.

The study compares the predictive performance of individual machine learning models, including XGBoost, Random Forest, Linear Regression, and Feedforward Neural Networks, with ensemble learning approaches based on stacking and blending. The dataset was divided into a model development subset and an unseen holdout sample. Model performance was evaluated using Mean Absolute Error, Root Mean Squared Error, and Symmetric Mean Absolute Percentage Error. The results indicate that XGBoost performed strongly among the individual models due to its ability to capture nonlinear relationships and feature interactions. Among the ensemble approaches, the blending model achieved slightly better predictive performance than stacking, suggesting its potential for improving prediction stability and generalizability.

In addition to predictive accuracy, the study examined feature importance to enhance model interpretability. Product availability, discount percentage, and pack size were identified as important factors influencing pharmaceutical price prediction. These findings suggest that machine learning can support more transparent and data-driven pricing decisions in developing markets. The study contributes to the pharmaceutical pricing literature by comparing individual and ensemble machine learning methods in a developing-market context and linking predictive modeling with practical pricing insights for companies and policymakers.

Keywords: pharmaceutical pricing; machine learning; ensemble learning; XGBoost; Random Forest; pharmaceutical market; Pakistan.

发展中市场中数据驱动的药品定价：一种机器学习方法

摘要： 发展中市场的药品定价往往受到透明度有限、可负担性约束、产品供应不均衡以及定价结构效率不足等因素的影响。这些问题使制药企业、医疗服务提供者和政策制定者难以作出基于证据的定价决策。本研究构建了一个用于预测巴基斯坦市场药品价格的机器学习框架，所采用的产品层面变量包括公司名称、包装规格、折扣前价格、折扣比例以及产品可获得性。数据集来源于 Kaggle 上公开可获得的数据集，共包含 1,630 条药品产品记录。在模型开发之前，本研究进行了探索性数据分析和数据预处理，以考察价格分布、高价值观测值、包装规格差异、缺失值以及类别特征。最终折扣后价格被设定为目标变量。

本研究比较了多种单一机器学习模型的预测性能，包括 XGBoost、随机森林、线性回归和前馈神经网络，并进一步比较了基于堆叠和混合的集成学习方法。数据集被划分为模型开发子集和最终未见过的留出样本。模型性能采用平均绝对误差、均方根误差和对称平均绝对百分比误差进行评估。结果表明，XGBoost 在单一模型中表现较强，这主要归因于其捕捉非线性关系和特征交互的能力。在集成学习方法中，混合模型的预测性能略优于堆叠模型，说明其在提高预测稳定性和泛化能力方面具有一定潜力。

除预测准确性外，本研究还分析了特征重要性，以增强模型结果的可解释性。研究发现，产品可获得性、折扣比例和包装规格是影响药品价格预测的重要产品层面因素。研究结果表明，机器学习能够支持发展中市场中更加透明和数据驱动的药品定价决策。本研究通过在发展中市场背景下比较单一机器学习模型与集成学习方法，并将预测建模与制药企业和政策制定者的实际定价洞察相结合，为药品定价研究文献作出了贡献。

关键词： 药品定价；价格预测；机器学习；集成学习；XGBoost；发展中市场；巴基斯坦。

1. Introduction

Pharmaceutical pricing is a critical issue in healthcare systems because it directly affects the accessibility, affordability, and availability of medicines. It also has broader implications for public health outcomes, market efficiency, and economic sustainability. In developing markets, pharmaceutical pricing is often characterized by limited transparency, regulatory inefficiencies, uneven market access, and price distortions caused by policy interventions such as

price controls and subsidies [4,6]. These conditions make it difficult to design pricing strategies that both protect consumers and maintain the financial sustainability of pharmaceutical firms. As a result, substantial variations in medicine prices are frequently observed, which may reduce consumer welfare and weaken the overall effectiveness of healthcare delivery systems. This situation highlights the need for more advanced and reliable pricing models capable of capturing the broad range of factors influencing pharmaceutical prices [4,6,10].

Traditionally, pharmaceutical prices have been shaped by production costs, market demand, competition, and regulatory frameworks. However, in developing economies, conventional pricing approaches often fail to account for several important determinants, such as packaging characteristics, discount structures, product availability, and market-specific conditions. The growing complexity of pharmaceutical markets has made traditional pricing approaches less effective, particularly in contexts where external economic pressures and policy changes frequently affect pricing decisions [4,6]. This has created a strong need for data-driven methods that can more effectively model pharmaceutical pricing behavior in such environments [1,5].

Machine learning (ML), as a major branch of artificial intelligence (AI), provides an attractive alternative for solving complex prediction problems by learning patterns directly from data rather than relying solely on predefined statistical assumptions. In pharmaceutical and healthcare research, AI and ML have increasingly been recognized for their ability to extract meaningful relationships from large and complex datasets [2,3,11]. In the context of pharmaceutical pricing, ML can improve prediction accuracy by incorporating multiple variables simultaneously, including product characteristics, market conditions, and policy-related factors. Because pharmaceutical pricing is influenced by nonlinear interactions among many variables, ML techniques are often more suitable than traditional linear approaches for this purpose [1,2,5].

The development of ML-based pricing models is particularly important in developing markets, where pharmaceutical companies often struggle to set optimal prices in the absence of efficient pricing systems. In many of these settings, patients rely heavily on out-of-pocket expenditure for medicines, making accurate pricing decisions essential for both affordability and profitability [4,10]. Conventional pricing methods, such as cost-plus pricing and competitor-based pricing, may not be flexible enough to reflect market volatility, product shortages, discount practices, and changes in regulatory policies. In contrast, data-driven pricing approaches can adapt more effectively to changing market conditions and support better pricing decisions [5,9].

Recent advances in data-driven modeling have further strengthened the use of ML in pharmaceutical and healthcare applications. Data-driven methods are now widely used in pharmaceutical processes, healthcare analytics, and pharmacological research because of their ability to manage high-dimensional data and uncover hidden patterns that traditional methods may overlook [1,2,5,11]. In pharmaceutical sciences, ML and deep learning tools have increasingly been applied to prediction, classification, optimization, and decision support, demonstrating their practical value in complex domains [3]. These developments suggest that similar approaches can be effectively extended to pharmaceutical price prediction, especially in settings where pricing is influenced by multiple interacting features.

Among ML methods, ensemble learning has emerged as one of the most effective strategies for improving predictive performance. Ensemble learning combines the outputs of multiple models to reduce error, improve robustness, and enhance generalization [8]. Random Forest, introduced by Breiman, is one of the most widely used ensemble algorithms and is known for its ability to handle high-dimensional data, reduce overfitting, and model nonlinear relationships effectively [7]. These properties make it highly suitable for pharmaceutical pricing problems, where price determination often depends on many interrelated product and market attributes.

In addition to Random Forest, more advanced predictive approaches, such as gradient boosting and neural networks, have become increasingly important in health-related analytics. Deep learning methods, including feedforward neural networks, are particularly useful for identifying complex nonlinear interactions within data and have shown strong performance across several pharmaceutical and biomedical prediction tasks [3,11,15]. Computational intelligence approaches have also been successfully applied in pharmaceutical research to model complex scientific relationships, such as solubility behavior and drug-related outcomes, further demonstrating the value of intelligent predictive systems in pharmaceutical decision-making [15,16]. These advances indicate that sophisticated ML models can offer substantial benefits for price prediction compared with traditional approaches.

Although individual ML models can achieve strong predictive performance, the literature increasingly shows that combining multiple models often produces better results than relying on a single model alone. Ensemble-based pipelines have been successfully used in healthcare diagnosis, fracture prediction, medical image analysis, and healthcare cost prediction, where they improved both prediction accuracy and generalizability [12,13,14,17]. These

findings are especially relevant for pharmaceutical pricing, where data may be noisy, incomplete, and influenced by heterogeneous product- and market-level factors. By combining models with different strengths, ensemble learning can provide more stable and accurate predictions than any single model used in isolation [8,12,13].

Two particularly important ensemble methods are stacking and blending. In stacking, several base models are first trained independently, and their outputs are then used as inputs for a meta-model that generates the final prediction [8,13,14]. This approach is effective when the base models capture different aspects of the data, allowing the meta-model to learn how to combine their predictions efficiently. Blending is closely related to stacking, but it usually relies on a holdout strategy or separate data subsets to combine the outputs of different models and improve generalization. Both approaches benefit from model diversity and have been shown to perform well in complex prediction tasks in healthcare and related analytical domains [12,13,14,17].

In pharmaceutical price prediction, stacking and blending are especially promising because they allow the integration of models with complementary strengths. For example, tree-based models such as Random Forest can capture nonlinear feature relationships and interactions, regression-based models offer interpretability, and neural network approaches can detect deeper hidden patterns in the data. The integration of these capabilities can generate a more powerful and reliable framework for pharmaceutical pricing than any standalone model. Given the increasing role of data-driven decision-making in pharmaceutical sciences and health economics, the use of ensemble ML methods provides a practical and theoretically grounded solution for addressing pricing challenges in developing markets [1,3,8,11].

The existing literature highlights the growing potential of machine learning to enhance pharmaceutical pricing decisions, particularly in developing economies where pricing systems often lack efficiency and transparency. Advances in artificial intelligence, deep learning, and ensemble methods in healthcare analytics demonstrate the ability of data-driven approaches to model complex pricing dynamics more accurately than traditional techniques [2,3,11,12,17]. Despite these developments, the application of predictive machine learning frameworks to pharmaceutical pricing, especially frameworks incorporating product-level characteristics, remains limited.

This study addresses this gap by developing and evaluating a comprehensive machine learning framework for pharmaceutical price prediction in a developing-market context. Its primary contribution lies in the systematic comparison of both individual models and advanced ensemble techniques, including stacking and blending, using product-specific variables such as availability, discount structures, and pack size. In contrast to prior studies that largely emphasize pricing policies, affordability, and macro-level determinants, this research adopts a micro-level predictive approach capable of capturing nonlinear relationships and complex interactions within pricing data.

By benchmarking models such as Linear Regression, Random Forest, XGBoost, and Feedforward Neural Networks against ensemble approaches, the study provides rigorous empirical evidence on model performance, robustness, and generalizability. This comparative framework not only identifies the most effective predictive techniques but also demonstrates the added value of ensemble learning in improving prediction accuracy.

The study is guided by the following objectives:

1. To develop a robust machine learning framework for pharmaceutical price prediction.
2. To evaluate and compare the performance of individual and ensemble models.
3. To quantify the impact of product-level features on pricing outcomes.
4. To generate actionable insights for transparent and competitive pricing strategies.

Overall, this work advances the pharmaceutical pricing literature by integrating health economics with modern machine learning methodologies. It introduces a novel data-driven predictive framework that extends beyond descriptive and policy-oriented analyses. By improving pricing accuracy, transparency, and adaptability, the proposed approach provides practical value for policymakers, regulators, and industry stakeholders and ultimately contributes to more efficient markets, improved affordability, and enhanced healthcare access in developing economies.

2. Methodology

This section presents the methodological framework adopted to predict pharmaceutical prices in the Pakistani market using machine learning techniques. The proposed framework follows a structured and reproducible pipeline that includes data acquisition, exploratory data analysis, preprocessing, feature engineering, model development, and performance evaluation. Both individual machine learning models and advanced ensemble approaches are implemented to capture complex nonlinear relationships in pricing data and to improve predictive accuracy and generalizability.

2.1 Data Description and Characteristics

The dataset used in this study was obtained from a publicly available Kaggle repository and contains 1,630 pharmaceutical product records from the Pakistani market. The dataset includes product-level attributes such as company name, pack size, pre-discount price, post-discount price, discount percentage, and product availability. The target variable in this study is the final

post-discount pharmaceutical price, while the remaining variables are treated as predictors.

To improve dataset transparency and reproducibility, an exploratory review of the dataset was conducted before model development. The original price variables were strongly right-skewed, indicating that most pharmaceutical products were concentrated at lower price levels, whereas a smaller number of products had substantially higher prices. Therefore, the distribution of the post-discount price was visualized using a log-transformed scale, log(1 + price), as shown in Figure 1. This transformation was used only for exploratory visualization, while the original price values were retained for model development.

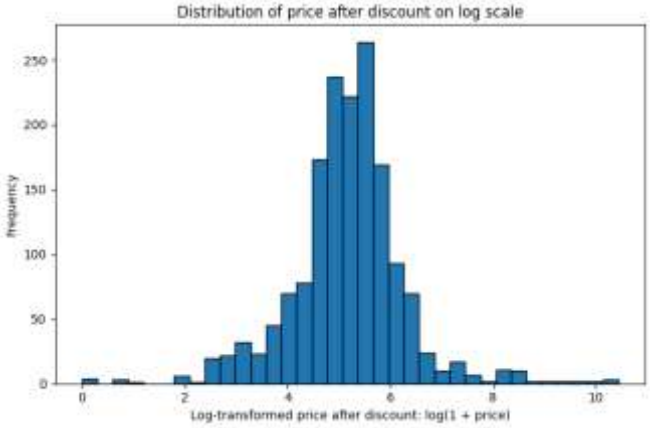

Figure 1. Distribution of price after discount on log scale. (Developed by Authors)

The boxplot in Figure 2 further shows the presence of high-value observations in the price-after-discount variable. These observations were retained because pharmaceutical pricing data may naturally include expensive products, and removing them could reduce the real-world variability of the dataset.

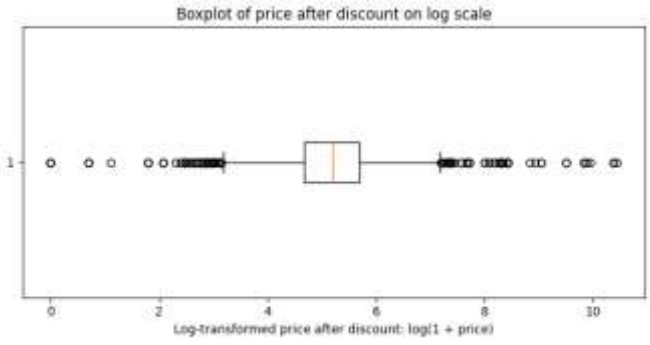

Figure 2. Boxplot of price after discount on log scale. (Developed by Authors)

Pack size was treated as a categorical feature and standardized before model development. Figure 3 shows the ten most frequent pack-size categories after standardization, indicating an uneven distribution of packaging formats across pharmaceutical products. This variability reflects differences in product quantity, dosage form, and packaging practices in the pharmaceutical market.

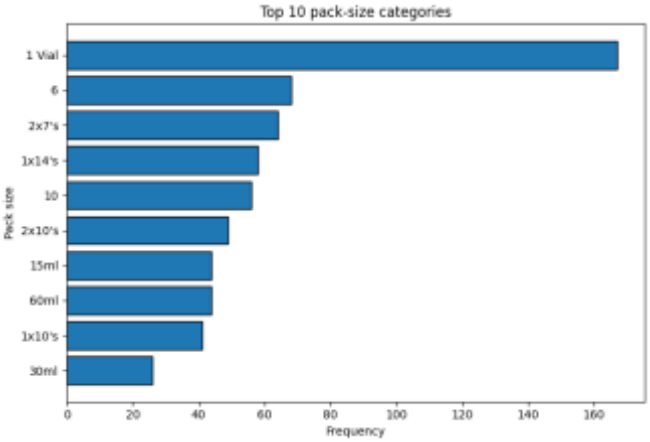

Figure 3. Top 10 pack-size categories after standardization. (Developed by Authors)

Discount percentage showed limited variation in the available records, with most recorded observations concentrated around similar discount levels. Availability was standardized during preprocessing and treated as a binary feature representing whether a product was available or unavailable in the market. Since the dataset was obtained from a public online source, some records required cleaning and standardization before model development.

It is important to acknowledge that potential dataset biases may exist due to the use of a publicly available online source. These may include overrepresentation of certain companies or product categories and possible inconsistencies caused by online data collection. Therefore, the findings should be interpreted within the scope of the available product-level data rather than as a complete representation of all pharmaceutical pricing dynamics in developing markets.

2.2 Data Preprocessing and Feature Engineering

Before model training, the dataset underwent a structured preprocessing and feature engineering pipeline to ensure compatibility with machine learning algorithms and to enhance reproducibility:

- **Missing Values:** Missing numerical values were handled using mean imputation via the *Simple Imputer* in scikit-learn. This approach preserves dataset size while maintaining central tendency.
- **Categorical Encoding:**
  - ✓ Company names were encoded using label encoding due to high cardinality.
  - ✓ Pack size was treated as a categorical variable and encoded numerically.

- ✓ Availability was encoded as a binary variable.

- **Feature Transformation:** The target variable was defined as the final price after discount. Discount percentage was retained as a continuous predictor to capture pricing adjustments.
- **Feature Scaling:** Continuous variables were normalized where required, particularly for the Feedforward Neural Network model, to improve convergence and training stability.

No synthetic or derived features were introduced; however, the selected variables inherently capture key pricing determinants, including supply conditions (availability), pricing strategy (discount), and product configuration (pack size).

2.3 Data Splitting Strategy

To ensure a rigorous evaluation framework and to prevent data leakage, the dataset was divided into two mutually exclusive subsets:

- **Model development dataset:** 98%
- **Final holdout dataset:** 2%

The model development dataset was evaluated using a 10-fold cross-validation procedure, in which the data were partitioned into ten subsets, with each subset used once for testing and the remaining subsets for training.

The final holdout dataset (2%) was excluded from this process and used solely for final prediction, enabling an unbiased assessment of model generalization under real-world conditions.

2.4 Machine Learning Algorithms

This section outlines the methodology used to develop the prediction models in this study. The primary objective is to assess the performance of four independent machine learning models XGBoost, Random Forest, Linear Regression, and Feedforward Neural Network (FNN) to identify the most effective approach for predicting the prices of pharmaceutical products. After evaluating each model individually, the study proceeds to combine the models using advanced ensemble learning techniques, specifically stacking and blending, with Linear Regression as the meta-model.

Phase 1: Individual Model Evaluation

In this phase, four distinct machine learning models are trained and evaluated independently to determine their effectiveness in predicting pharmaceutical prices. Each model is tested using a subset of the 98% of the dataset, with the 2% held out for final prediction.

The models selected are:

a) XGBoost (Extreme Gradient Boosting)

XGBoost is an optimized gradient-boosting algorithm that builds an ensemble of weak learners, typically decision trees, to improve model accuracy. XGBoost minimizes a regularized objective function consisting of a loss function and a complexity term to reduce overfitting.

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

where $L(y_i, \hat{y}_i)$ is the loss function, such as Mean Squared Error (MSE), between the actual y_i and predicted $\hat{y}_i$, $\Omega(f_k)$ is the regularization term for the complexity of each treef_k, The model prediction is calculated by summing the output of all trees

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

where $f_k(x_i)$ is the prediction from the k-th $tree$ for the i-th sample.

XGBoost is highly effective for price prediction as it handles large datasets and complex interactions between variables, providing robust predictions with low error rates.

b) Random Forest

Random Forest is an ensemble of decision trees, where each tree is built from a random subset of features and samples.

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x)$$

Where $T_b(x)$ is the prediction from the b-th tree, B is the total number of trees.

Random Forest uses "bagging" (Bootstrap aggregating) by training each tree on a bootstrap sample from the data, reducing variance and overfitting. It is suitable for handling large datasets with complex relationships and is robust against overfitting, providing stable predictions even in the presence of noisy data.

c) Linear Regression:

Linear Regression is a simple regression technique that assumes a linear relationship between the independent variables and the dependent variable. The model predicts the output using a linear combination of input feature.

$$\hat{y} = \beta_0 + \sum_{j=1}^{\Lambda} \beta_j x_j$$

where $\hat{y}$ is the predicted output, β_0 is the intercept, β_j is the coefficient for feature j, x_j is the value of feature$j$. Linear Regression minimizes the Mean Squared Error (MSE) between the actual and predicted values to find the best-fit line.

$$\mathrm{MSE} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Although simpler than other models, it can serve as a baseline, especially for interpretable insights about linear relationships in the data.

d) Feedforward Neural Network (FNN)

The Feedforward Neural Network is a type of artificial neural network (ANN) where connections between the nodes do not form cycles. It consists of multiple layers: an input layer, hidden layers, and an output layer. The model passes inputs through multiple layers of neurons with activation functions to transform inputs into outputs.

$\hat{y} = \sigma(W_n\sigma(W_{n-1}\dots\sigma(W_1x + b_1) + b_{n-1}) + b_n)$

where x is the input vector, W_i and b_i represent the weights and biases for layer i, σ is an activation function, commonly ReLU (Rectified Linear Unit) or Sigmoid for FNNs.

The network uses backpropagation to minimize the loss function, typically MSE for regression, by adjusting weights through gradient descent.

$$\mathrm{Loss} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

FNNs capture complex nonlinear relationships in data, making them useful for pharmaceutical price prediction where interactions between features like discount, pack size, and availability can influence prices.

Phase 2: Ensemble Methods: Stacking and Blending

After evaluating each of the individual models, we apply both **stacking** and **blending** ensemble methods to improve prediction accuracy by combining the strengths of multiple models.

a) Stacking ensemble model

Stacking (or stacked generalization) is a technique where predictions from multiple base models (XGBoost, Random Forest, Linear Regression, FNN) are used as input features for a meta-model. The meta-model, often a simpler model like Linear Regression, is trained on the predictions of the base models, combining their outputs to make the final prediction. This approach can capture complex relationships between the base model predictions, allowing for improved performance.

i) Base Learners (Level-0 Models)

The stacking ensemble begins by training multiple base models (referred to as level-0 models) on the same dataset. Given a dataset$D = \{(x_i, y_i)\}_{i=1}^{n}$, where x_i represents the feature vector and y_i is the target variable, the predictions of each model are as follows:

For a base model f_j :

$$\hat{y}_{i,j} = f_j(x_i),\ j = 1, \dots, M$$

where M represents the number of base models, and $\hat{y}_{i,j}$ denotes the prediction of the j-th model for the i-th instance.

The output of each base model f_j serves as a new feature for the next stage.

ii) Meta-Learner (Level-1 Model)

A meta-learner, or level-1 model, takes the predictions of the base models as inputs and learns to make a final prediction. The input for the meta-learner is the prediction matrix$Z$, defined as:

$$Z = \begin{matrix} \hat{y}_{1,1} & \hat{y}_{1,2} & \cdots & \hat{y}_{1,M} \\ \hat{y}_{2,1} & \hat{y}_{2,2} & \cdots & \hat{y}_{2,M} \\ \vdots & \vdots & \ddots & \vdots \\ \lfloor\hat{y}_{n,1} & \downarrow\ un,2 & \cdots & \hat{y}_{n,M}\rfloor \end{matrix}$$

The meta-learner g then learns from Z and outputs the final prediction $\hat{y}$:

$$\hat{y} = g(Z)$$

iii) Mathematical Formulation

For a new instancex_{new}, the final prediction $\hat{y}_{\mathrm{new}}$ from the stacking ensemble is obtained by first generating predictions from each base learner and then applying the meta-learner on these predictions:

$$\hat{y}_{new} = g\big(f_1(x_{new}), f_2(x_{new}), \dots, f_M(x_{\mathrm{new}})\big)$$

b) Blending ensemble model

Blending is a similar ensemble method but differs in how the data is split. Instead of cross-validation, blending uses a hold-out validation set (part of the data reserved for testing). The base models make predictions on the validation set, and a meta-model is trained using these predictions to generate the final output. Blending is often simpler and faster to implement than stacking, as it requires less computational effort and avoids the complexity of cross-validation.

Blending is similar to stacking but uses a holdout set to train the meta-learner instead of using cross-validation across the entire dataset.

The dataset is split into a training set for model development and a 10–20% holdout set for validation. The predictions on this holdout set are used as features for the meta-learner.

i) Base Learner Training

Each base model f_j is trained on the training set and then used to predict the outcomes on the holdout set:

$$\hat{y}_{i,j} = f_j(x_i), \text{ for } x_i \in \text{ Holdout Set}$$

$\hat{y}_{i,j} = f_j(x_i)$, for $x_i \in$ HoldoutSet

i) Meta-Learner Training and Prediction

The meta-learner g is then trained on the predictions from the holdout set and makes a final prediction as in stacking.

For a new instance x_{new} :

$$\hat{y}_{new} = g\left(f_1(x_{new}), f_2(x_{new}), \ldots, f_M(x_{new})\right)$$

Based on the evaluation metrics (MSE, MAE, RMSE, SMAPE), we found that blending provided superior performance in terms of predictive accuracy compared to stacking. As a result, the blending ensemble method was selected for the final prediction.

By using the blending ensemble, we achieved a more robust and accurate model for pharmaceutical price prediction, capturing a broader range of patterns in the data and improving generalization on unseen data.

2.5 Evaluation Metrics for Model Performance

The following performance evaluation metrices are used to compare the performance of the models applied to the pharmaceutical data.

2.4.1. Mean absolute error (MAE)

MAE measures the average absolute difference between the actual and predicted values. Unlike MSE, MAE doesn't square the differences, so it provides a more interpretable error measure that treats all errors equally.

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

where $|y_i - \hat{y}_i|$ is absolute difference between the actual and predicted values.

MAE provides an easier-to-understand measure of error because it's in the same units as the target variable. A lower MAE indicates a better model.

2.4.2 Root mean square error (RMSE)

RMSE is the square root of the Mean Squared Error (MSE). It provides an error metric that penalizes larger errors more heavily, similar to MSE, but it is in the same units as the target variable, making it more interpretable.

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

RMSE is useful for understanding how much error the model is introducing in terms of the original data units. Smaller RMSE values indicate better performance.

2.4.3 Symmetric mean absolute percentage error (SMAPE)

SMAPE is a percentage-based error metric that treats over-predictions and under-predictions equally, giving a more balanced view of prediction performance, especially when the data has both very small and large values.

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{(|y_i| + |\hat{y}_i|)/2}$$

SMAPE gives an error percentage that is normalized by the average of the actual and predicted values. It ranges from 0% to 200%, with 0% being perfect prediction. It's particularly useful when comparing models across datasets with different scales.

2.6 Model Training and Final Prediction

Model training and evaluation were conducted using a 10-fold cross-validation framework applied to the model development dataset. In each iteration, the model was trained on nine folds (90% of the data) and tested on the remaining fold (10%), ensuring that each observation served as validation data exactly once. This approach provides a reliable and stable assessment of model performance while effectively utilizing the available data.

Model performance was evaluated using MAE, RMSE, and SMAPE. The final performance metrics were obtained by averaging the results across all ten folds, thereby reducing variability associated with any single partition and yielding a robust estimate of predictive accuracy.

The final 2% holdout dataset was entirely excluded from the cross-validation process and was not used during model training or validation. After completing the cross-validation procedure, the trained model was applied to this holdout set solely for final prediction. This design ensures an unbiased evaluation of model generalization by simulating real-world

deployment conditions, where predictions are generated for completely unseen data.

3. Results and Discussions

The results obtained using different models are reported and discussed below.

3.1 Performance Evaluation of Individual Machine Learning Models

Table 1 presents the comparative performance of XGBoost, Feedforward Neural Network, Random Forest, and Linear Regression using MAE, RMSE, and SMAPE. The results show that XGBoost consistently achieves the lowest MAE (0.8039) and SMAPE (0.53%), indicating superior predictive accuracy and stability across different error measures. This strong performance can be attributed to its boosting framework, which sequentially minimizes residual errors and effectively captures nonlinear relationships and complex feature interactions. In addition, built-in regularization mechanisms help control overfitting and improve generalization.

Table 1. Performance metrices using Test Data (Compiled by Authors)

Model	MAE	RMSE	SMAPE
XGBoost	0.8039	2.4559	0.53%
Feedforward NN	1.7618	2.5993	2.22%
Random Forest	2.4107	15.3083	0.77%
Linear Regression	1.3599	2.1764	1.87%

The Feedforward Neural Network demonstrates comparatively higher error values (MAE = 1.7618, SMAPE = 2.22%), suggesting that the model may not have fully leveraged the underlying data structure. Neural networks typically require larger datasets and extensive hyperparameter tuning; in moderate-sized structured datasets, they may underperform relative to tree-based ensemble methods.

The Random Forest model exhibits a substantially higher RMSE (15.3083), despite maintaining a relatively low SMAPE (0.77%). This pattern indicates the presence of large prediction errors for certain observations, as RMSE is particularly sensitive to extreme deviations. The averaging nature of Random Forest can lead to oversmoothing, limiting its ability to capture sharp variations or outliers in the response variable.

In contrast, Linear Regression shows competitive performance in terms of RMSE (2.1764), even outperforming XGBoost on this metric. This suggests that the relationship between predictors and the response variable contains a meaningful linear component. However, its higher SMAPE (1.87%) reflects limitations in handling relative errors and capturing nonlinear patterns.

Overall, the results indicate that XGBoost provides the most balanced and reliable performance across multiple evaluation metrics, making it well-suited for the given dataset. The consistency of XGBoost's superior performance across multiple evaluation metrics suggests its robustness, although formal statistical significance testing was not conducted in this study. While Linear Regression remains competitive under partial linearity, and Random Forest and Neural Networks offer flexibility, their performance is comparatively less stable under the current data conditions.

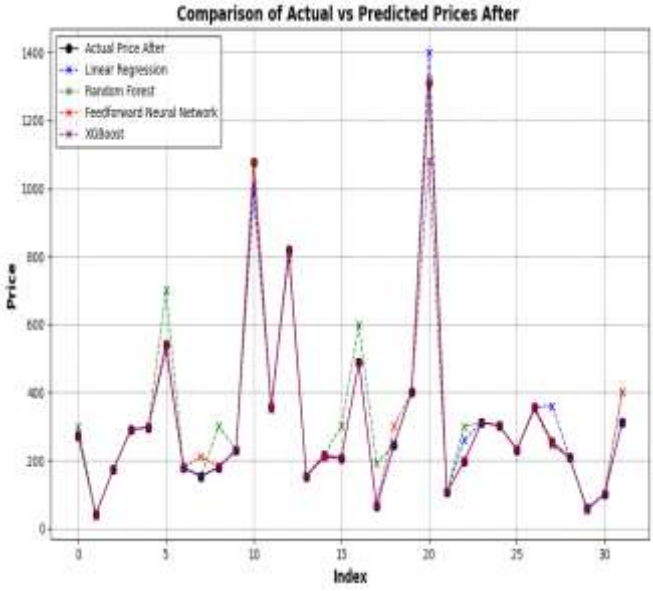

Figure 4: Comparison of actual and predicted price after of different models (Developed by Authors)

The Figure 4 shows XGBoost closely matching actual prices, demonstrating superior accuracy. Feedforward Neural Network (red) performs moderately well, with slight deviations. Linear Regression approximates trends but struggles with higher prices.

3.1.1 Model Interpretability and Feature Impact on Pharmaceutical Pricing

Model interpretability is important in pharmaceutical pricing because pricing decisions affect affordability, market access, and regulatory transparency. A model that only predicts prices accurately may have limited practical value if the main drivers of prediction are not understood. Therefore, this study examined feature importance across the applied models to identify the relative influence of availability, discount, and pack size on pharmaceutical pricing. This analysis uses multiple models Linear Regression, Random Forest, XGBoost, and Feedforward Neural Network to assess the Random Forest shows significant

inaccuracies, especially at extremes. Overall, XGBoost outperforms the other models in predictive reliability.

Importance of three key product features (Discount, Pack Size, and Availability) on pharmaceutical pricing. The normalized importance of each feature, as shown in the bar chart, provides insights into how strongly each factor influences pricing across different modeling techniques.

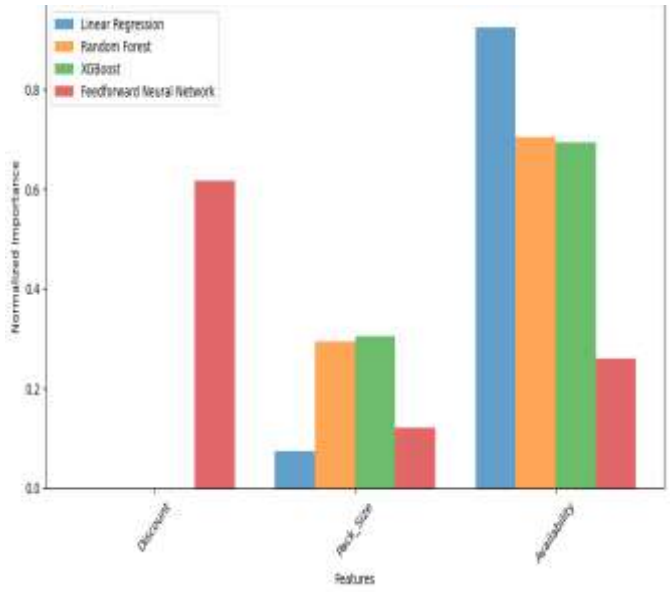

Figure 5: Normalized Feature Importance of Different Models (Developed by Authors)

Availability is the most influential feature across models, especially in Linear Regression, Random Forest, and XGBoost, with Linear Regression attributing the highest importance to it. This suggests that product availability is a key factor in determining pricing, and pharmaceutical companies should focus on optimizing supply chains to ensure high availability, potentially allowing for premium pricing.

Discount plays a significant role in the Feedforward Neural Network model, highlighting its complex impact on pricing, especially in competitive or price-sensitive markets. This implies that pharmaceutical companies should tailor discount strategies to specific customer segments or regions to optimize pricing outcomes.

Pack Size has a moderate influence in Random Forest and XGBoost, indicating it affects pricing but is less important than Availability and Discount. Companies may adjust pack sizes to improve affordability or offer premium products, but this should be secondary to availability and discount strategies.

Linear Regression emphasizes Availability, suggesting a simpler, linear relationship, while Random Forest and XGBoost capture more complex interactions, balancing Availability and Pack Size. The Feedforward Neural Network prioritizes Discount, reflecting its ability to model nonlinear patterns in demand and pricing.

The interpretability results show that availability, discount, and pack size are not only predictive variables but also practically meaningful pricing determinants. Availability reflects supply-side conditions and can influence market price variation, especially when products are scarce or inconsistently supplied. Discount captures pricing adjustments and competitive strategies, while pack size reflects product quantity and affordability considerations. These findings make the model more useful for decision-makers because they explain which product-level factors contribute most to price prediction. Although SHAP-based interpretation was not included in the present study, future research may apply SHAP values to provide observation-level explanations of individual price predictions.

3.2 Evaluation of Ensemble Models

After evaluating individual models, the next step is to analyze the performance of ensemble methods, specifically stacking and blending, to determine which provides the best predictive performance for pharmaceutical price prediction. Ensemble methods, which combine multiple individual models, are known for enhancing prediction accuracy and reducing model variance. These approaches are critical in addressing the second research objective of comparing the predictive performance of machine learning algorithms and ensemble techniques to identify the most robust and accurate model.

Table 2. Performance metrices using Test Data (Compiled by Authors)

Model	MAE	RMSE	SMAPE
Blending	1.1548	2.1207	1.86%
Stacking	1.3628	2.1868	1.88%

The Blending Model demonstrates strong performance on the test dataset, achieving an MAE of 1.1548, an RMSE of 2.1207, and a SMAPE of 1.86%. These metrics reflect its accuracy and reliability, offering a good balance between predictive accuracy and generalization. Its performance is comparable to XGBoost, the best-performing individual model, highlighting its potential for pharmaceutical price prediction tasks. The Stacking Model also performs well, with an MAE of 1.3628, an RMSE of 2.1868, and a SMAPE of 1.88%. Although slightly less accurate than the Blending Model, its consistent performance suggests effective generalization and minimal overfitting, making it a reliable choice for unseen data. Both ensemble methods provide robust results, with the Blending Model slightly outperforming the Stacking

Model in accuracy. While the Blending Model may be the better option for precise predictions, the Stacking Model remains a solid alternative with consistent performance.

The difference between the blending and stacking models was relatively small, indicating that both ensemble techniques provided stable performance. However, the blending model achieved lower MAE and RMSE than the stacking model, suggesting slightly better average prediction accuracy on the testing data. This may be because blending uses a holdout-based strategy for training the meta-model, which can reduce overfitting and improve generalization when the dataset is limited in size. Stacking also performed well, but its effectiveness depends on how base-model predictions are generated and combined by the meta-learner.

The ensemble results show that combining multiple models can improve prediction stability by using the complementary strengths of different algorithms. Tree-based models such as XGBoost and Random Forest can capture nonlinear interactions, Linear Regression provides a simple and interpretable baseline, and Feedforward Neural Networks can model more complex nonlinear patterns. By combining these models, the blending and stacking approaches produced robust predictive results for pharmaceutical price estimation.

3.3 Prediction using the Blending Ensemble Model

The Blending model will predict a random 2% sample of the original dataset. Chosen for its strong performance and ability to minimize overfitting, it combines base models with Linear Regression as the meta-model for balanced accuracy and generalization. This approach will deliver actionable insights, supporting strategic pharmaceutical pricing decisions in developing markets.

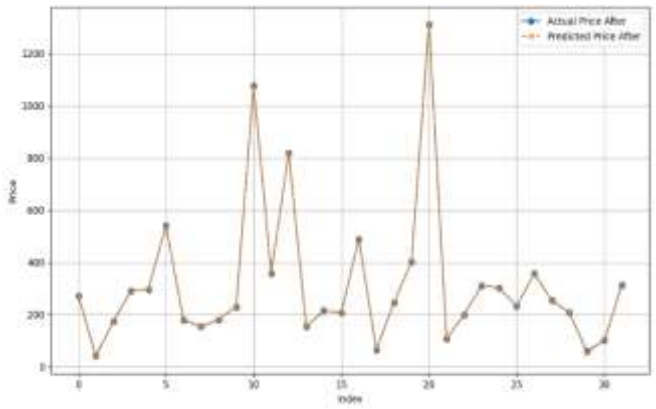

Figure 6: Comparison of Actual vs Predicted Prices After (Developed by Authors)

The figure 6 compares actual and predicted "Price After" values for pharmaceutical products using a blending ensemble model. The blue line represents actual prices, while the orange line shows predicted prices. The close alignment between the lines indicates high accuracy, with the model capturing price fluctuations, including larger spikes. Minor deviations suggest low error metrics, such as MAE and RMSE. Overall, the graph demonstrates the blending ensemble model's strong predictive accuracy, confirming its reliability for pharmaceutical pricing predictions.

4. Practical Implications for Pharmaceutical Pricing Strategies in Developing Markets

The findings of this study provide several practical implications for improving pharmaceutical pricing strategies in developing markets. One important implication is the role of product availability as a predictor of pharmaceutical price, particularly in the XGBoost model. This suggests that consistent product availability is closely related to pricing behavior and should be considered when developing pricing strategies. For pharmaceutical companies, availability information can support supply-chain planning and price forecasting. For regulators, it can help monitor whether price changes are associated with supply shortages or market instability.

Another important implication concerns discount strategies. The results indicate that discounts have a meaningful influence on pharmaceutical prices, especially in the Feedforward Neural Network model. This suggests that discount-price relationships may be nonlinear and may vary across products or market conditions. Pharmaceutical firms can use this information to design more targeted discount policies that reflect local demand conditions, consumer purchasing power, and price sensitivity. Such strategies may help improve medicine affordability while maintaining market competitiveness.

The analysis also shows that pack size is an important product-level factor, particularly in the XGBoost and Random Forest models. This finding suggests that firms can use pack-size decisions as part of their pricing strategies. Smaller pack sizes may improve affordability for low-income consumers, while larger pack sizes may provide better value for consumers who can purchase medicines in greater quantities. Therefore, pack-size planning can help pharmaceutical companies serve different consumer segments more effectively in developing markets.

In addition, the study demonstrates that ensemble methods, especially blending and stacking, can support more reliable pharmaceutical price prediction by combining the strengths of multiple models. This has practical value for both firms and policymakers. Pharmaceutical companies can use such models to generate data-driven price forecasts and reduce uncertainty in pricing decisions. Policymakers and regulators can use model-based evidence to support

more transparent monitoring of medicine prices and affordability. In this way, machine learning models can provide decision-support tools rather than replace expert judgment.

Finally, the study highlights the broader importance of interpretable and transparent pricing models in developing markets. A data-driven approach allows pharmaceutical prices to be examined using measurable factors such as availability, discount percentage, and pack size. This can help companies explain pricing decisions more clearly and assist policymakers in evaluating whether prices remain aligned with public health objectives. Therefore, machine learning-based pricing models can contribute not only to business efficiency but also to improved access to essential medicines, fairer pricing practices, and greater market transparency.

5. Conclusion and Future Work

This study developed a machine learning framework for predicting pharmaceutical prices in developing markets using product-level data from the Pakistani pharmaceutical market. The study compared individual machine learning models, including XGBoost, Random Forest, Linear Regression, and Feedforward Neural Networks (FNN), as well as ensemble approaches based on stacking and blending. Before model development, exploratory data analysis and preprocessing were conducted to examine price distribution, high-value observations, pack-size variability, missing values, and categorical features. The findings show that machine learning methods can provide useful predictive support for pharmaceutical pricing decisions, especially in markets where pricing transparency, affordability, and availability remain major challenges.

Among the individual models, XGBoost demonstrated strong predictive performance because of its ability to capture nonlinear relationships and interactions among pricing-related variables. The ensemble results showed that both stacking and blending produced stable predictions, with the blending model achieving slightly better performance based on MAE, RMSE, and SMAPE. These results suggest that combining multiple models can improve prediction reliability and support more robust pharmaceutical price estimation.

The study also highlighted the importance of model interpretability for healthcare and pricing decisions. Feature-importance analysis showed that availability, discount percentage, and pack size were important product-level factors influencing pharmaceutical price prediction. These findings provide practical insights for pharmaceutical companies and policymakers. For companies, the results may support more informed decisions regarding discount strategies, pack-size planning, and supply management. For policymakers and regulators, the findings may help improve pricing transparency and support evidence-based monitoring of medicine affordability.

Despite these contributions, the study has some limitations. The dataset was obtained from a publicly available Kaggle source and represents available product-level information from the Pakistani pharmaceutical market. Therefore, the results may not fully represent all medicines, companies, regions, or informal pricing variations in developing markets. In addition, the model comparison was based on predictive error metrics rather than formal statistical significance testing. Although a final unseen holdout sample was used for prediction, future studies could further strengthen model validation by applying k-fold cross-validation or repeated random subsampling.

Future research may expand the proposed framework by using larger and more diverse pharmaceutical pricing datasets from multiple developing countries. Additional variables, such as market competition, therapeutic class, regulatory policy, inflation, exchange rates, procurement channels, and regional demand conditions, could also be included to improve prediction accuracy and practical relevance. Future work may also apply SHAP values or other explainable artificial intelligence methods to provide observation-level explanations of individual price predictions. These extensions would strengthen the reliability, interpretability, and policy usefulness of machine learning-based pharmaceutical pricing models.

Declarations

Author Contributions: Conceptualization, A.A.H.F., F.H.F., and Z.A.; methodology, A.A.H.F. and Z.A.; software, A.A.H.F.; validation, A.A.H.F., F.H.F., Z.A., S.A., and W.H.S.; formal analysis, A.A.H.F.; investigation, A.A.H.F.; resources, Z.A., S.A., and W.H.S.; data curation, A.A.H.F.; writing original draft preparation, A.A.H.F.; writing review and editing, F.H.F., Z.A., S.A., and W.H.S.; visualization, A.A.H.F.; supervision, Z.A.; project administration, F.H.F. and Z.A. All authors have read and agreed to the published version of the manuscript.

Data Availability Statement: The dataset supporting the findings of this study is publicly available on Kaggle at:
https://www.kaggle.com/datasets/talhasattar727/pakistan-pharmaceutical-dataset.

Acknowledgments: The authors would like to acknowledge the academic support and research environment provided by their respective institutions during the completion of this study.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest. In addition, the authors confirm that all ethical issues, including plagiarism, misconduct, data fabrication and/or falsification, double publication and/or submission, and redundancy, have been completely observed.

References

[1] Dong Y, Yang T, Xing Y, Du J, Meng Q (2023) Data-driven modeling methods and techniques for pharmaceutical processes. *Processes* 11(7):2096.

[2] Bhattamisra SK, Banerjee P, Gupta P, Mayuren J, Patra S, Candasamy M (2023) Artificial intelligence in pharmaceutical and healthcare research. *Big Data Cogn Comput* 7(1):10.

[3] Javid S, Rahmanulla A, Ahmed MG, Sultana R, Prashantha Kumar BR (2025) Machine learning & deep learning tools in pharmaceutical sciences: A comprehensive review. *Intelligent Pharmacy* 3(3):167-180. https://doi.org/10.1016/j.ipha.2024.11.003

[4] Abdel Rida N, Ibrahim MIM, Babar ZUD, Owusu Y (2017) A systematic review of pharmaceutical pricing policies in developing countries. *J Pharm Health Serv Res* 8(4):213-226. https://doi.org/10.1111/jphs.12191

[5] Dean EB (2019) Who benefits from pharmaceutical price controls? Evidence from India. *CGD Working Paper* 509. Center for Global Development, Washington, DC.

[6] Danzon PM, Chao LW (2000) Cross-national price differences for pharmaceuticals: How large, and why? *J Health Econ* 19(2):159-195.

[7] Bate R, Jin GZ, Mathur A (2011) Does price reveal poor-quality drugs? Evidence from 17 countries. *J Health Econ* 30(6):1150-1163.

[8] Santos MAB, Dias LLS, Pinto CDBS, Silva RM, Osorio-de-Castro CGS (2019) Factors influencing pharmaceutical pricing: A scoping review of academic literature in health science. *J Pharm Policy Pract* 12:24. https://doi.org/10.1186/s40545-019-0183-0

[9] Janssen Daalen JM, den Ambtman A, van Houdenhoven M, van den Bemt BJF (2021) Determinants of drug prices: A systematic review of comparison studies. *BMJ Open* 11(7):e046917. https://doi.org/10.1136/bmjopen-2020-046917

[10] Lee KS, Kassab YW, Taha NA, Zainal ZA (2021) A systematic review of pharmaceutical price mark-up practice and its implementation. *Explor Res Clin Soc Pharm* 2:100020.

[11] Lee KS, Kassab YW, Taha NA, Zainal ZA (2021) Factors impacting pharmaceutical prices and affordability: Narrative review. *Pharmacy* 9(1):1.

[12] Breiman L (2001) Random forests. *Mach Learn* 45(1):5-32.

[13] Polikar R (2012) Ensemble learning. In: Zhang C, Ma Y (eds) *Ensemble Machine Learning: Methods and Applications*. Springer US, New York, pp 1-34. https://doi.org/10.1007/978-1-4419-9326-7_1

[14] Wolpert DH (1992) Stacked generalization. *Neural Netw* 5(2):241-259. https://doi.org/10.1016/S0893-6080(05)80023-1

[15] Breiman L (1996) Stacked regressions. *Mach Learn* 24(1):49-64.

[16] Chen T, Guestrin C (2016) XGBoost: A scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 785-794.

[17] Kumar M, Nguyen TPN, Kaur J, Singh TG, Soni D, Singh R, Kumar P (2023) Opportunities and challenges in application of artificial intelligence in pharmacology. *Pharmacol Rep* 75(1):3-18.

参考文:

[1] Dong Y, Yang T, Xing Y, Du J, Meng Q (2023) 药物制药过程中的数据驱动建模方法与技术。*Processes* 11(7):2096.

[2] Bhattamisra SK, Banerjee P, Gupta P, Mayuren J, Patra S, Candasamy M (2023) 人工智能在制药与医疗保健研究中的应用。*Big Data Cogn Comput* 7(1):10.

[3] Javid S, Rahmanulla A, Ahmed MG, Sultana R, Prashantha Kumar BR (2025) 制药科学中的机器学习与深度学习工具：一项综合综述。*Intelligent Pharmacy* 3(3):167-180. https://doi.org/10.1016/j.ipha.2024.11.003

[4] Abdel Rida N, Ibrahim MIM, Babar ZUD, Owusu Y (2017) 发展中国家药品定价政策的系统综述。*J Pharm Health Serv Res* 8(4):213-226. https://doi.org/10.1111/jphs.12191

[5] Dean EB (2019) 谁从药品价格管制中受益？来自印度的证据。*CGD Working Paper* 509. Center for Global Development, Washington, DC.

[6] Danzon PM, Chao LW (2000) 药品跨国价格差异：差异有多大，以及原因何在？ *J Health Econ* 19(2):159-195.

[7] Bate R, Jin GZ, Mathur A (2011) 价格是否揭示劣质药品？来自17个国家的证据。*J Health Econ* 30(6):1150-1163.

[8] Santos MAB, Dias LLS, Pinto CDBS, Silva RM, Osorio-de-Castro CGS (2019) 影响药品定价的因素：健康科学学术文献的范围综述。*J Pharm Policy Pract* 12:24. https://doi.org/10.1186/s40545-019-0183-0

[9] Janssen Daalen JM, den Ambtman A, van Houdenhoven M, van den Bemt BJF (2021) 药品价格决定因素：比较研究的系统综述。*BMJ Open*

11(7):e046917. https://doi.org/10.1136/bmjopen-2020-046917
[10] Lee KS, Kassab YW, Taha NA, Zainal ZA (2021) 药品价格加成做法及其实施的系统综述。*Explor Res Clin Soc Pharm* 2:100020.
[11] Lee KS, Kassab YW, Taha NA, Zainal ZA (2021) **影响**药品价格与可负担性的因素：叙述性综述。*Pharmacy* 9(1):1.
[12] Breiman L (2001) **随机森林**。*Mach Learn* 45(1):5-32.
[13] Polikar R (2012) **集成学**习。载于：Zhang C, Ma Y（编）*Ensemble Machine Learning: Methods and Applications*。Springer US, New York, pp 1-34. https://doi.org/10.1007/978-1-4419-9326-7_1
[14] Wolpert DH (1992) **堆叠泛化**。*Neural Netw* 5(2):241-259. https://doi.org/10.1016/S0893-6080(05)80023-1
[15] Breiman L (1996) **堆叠回**归。*Mach Learn* 24(1):49-64.
[16] Chen T, Guestrin C (2016) XGBoost：**一种可**扩展的树提升系统。载于：*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 785-794.
[17] Kumar M, Nguyen TPN, Kaur J, Singh TG, Soni D, Singh R, Kumar P (2023) **人工智能在**药理学应用中的机遇与挑战。*Pharmacol Rep* 75(1):3-18.

Manuscript Information
Word count: 8,229 words (excluding references).
Peer-Review Record

Fast-track status: Not fast-tracked.
First-round reviews received: 3 reports.
Revision cycles completed: 3 rounds.
Final version submitted: June 5, 2026