Journal of Global Research in Electronics and Communication
Volume 2, No. 05, May 2026
Available Online at: www.jgrec.info
RESEARCH PAPER

Artificial Intelligence- Driven Analysis for Pharmaceutical Product Pricing and Composition

Dr. Parth Gautam
Associate Professor
Department of Computer Sciences and Applications
Mandsaur University, Mandsaur
parth.gautam@meu.edu.in

***Abstract*—Increased costs in the health sector have placed significant pressure on public budgets allocated for pharmaceutical procurement. In the context of financial crises and global health emergencies, national purchasing authorities face a critical challenge of obtaining high-quality pharmaceuticals at minimal cost. Although existing literature has explored various influencing factors using producer and reference price data, limited attention has been given to individual-level purchase data from diverse public buyers. To fill this gap, this study suggests an ensemble-based AdaBoost classification method using the DrugBank database which contains about 4900 pharmaceutical compounds. The dataset is subjected to robust preprocessing, such as handling missing values, removing duplicates, implementing One-Hot Encoding, normalizing data using the Z-score, and balancing the classes using RandomOverSampler, ensuring the quality and robustness of the data for modeling. To perform comparative analysis the following multiple machine learning models are evaluated including KNN, SVM, CNN, XGBoost and the proposed AdaBoost Classifier. Experimental results show that the proposed AdaBoost classifier has excellent predictive ability and can achieve an accuracy of 99%, a precision of 99%, a recall of 99%, and an F1 score of 99%. The results show the effectiveness of ensemble learning for achieving relationships in pharmaceutical data. The study concludes that the proposed model can be used as a solution with very high accuracy and reliability in predicting the price and composition of pharmaceutical products.**

Keywords—Drug pricing, AI-driven drug discovery, value-based pricing, cost-effectiveness, pharmaceutical R&D, healthcare economics.

I. INTRODUCTION

The pharmaceutical industry plays a crucial role in global healthcare by ensuring the availability of safe, effective, and affordable medicines. However, the increasing complexity of pharmaceutical products, coupled with rising medication costs, has created significant challenges for healthcare providers, policymakers, and consumers[1][2]. Accurate analysis of pharmaceutical product pricing and composition is essential for improving accessibility, maintaining quality standards, and supporting informed healthcare decisions. Traditional methods of pharmaceutical analysis often involve extensive manual effort, making them time-consuming, expensive, and susceptible to human errors [3].

The rapid growth of digital healthcare platforms and online pharmacies has generated large volumes of pharmaceutical data related to drug composition, pricing, availability, and consumer demand[4][5]. Despite the availability of such information, identifying pricing patterns and understanding the relationships between drug compositions and market value remain challenging tasks. Furthermore, increasing drug prices have become a major concern worldwide, affecting medication adherence and overall healthcare outcomes. Consequently, there is a growing need for intelligent data-driven approaches capable of analyzing pharmaceutical products efficiently and accurately [6][7].

Recent advancements in Artificial Intelligence (AI) have significantly transformed pharmaceutical research and development[8][9][10]. AI technologies facilitate the analysis of large-scale datasets, enabling faster and more reliable predictions regarding drug properties, toxicity, efficacy, and market trends. In particular, AI-driven analytical systems can support pharmaceutical pricing evaluation and composition assessment by discovering hidden patterns within complex datasets[11][12][13]. These capabilities can help healthcare organizations optimize pricing strategies, improve transparency, and enhance the overall efficiency of pharmaceutical management processes [14][15][16].

Among AI techniques, Machine Learning (ML) and Deep Learning (DL) have emerged as powerful tools for predictive analytics and decision-making in pharmaceutical applications[17]. ML algorithms can learn complex relationships between drug attributes and pricing information, while DL models are capable of extracting high-level representations from large and heterogeneous datasets[18][19][20][21].

A. Motivation and Contribution

The motivation behind this study the rising cost of pharmaceutical products has created a critical need for intelligent and data-driven pricing analysis systems. Large and complex biomedical data often require traditional analytical approaches that are inefficient. This can encourage the adoption of machine learning methods for uncovering patterns in drug composition and pricing. The objective of the research is to build an accurate and reliable predictive model that can aid in making healthcare decisions. This research offers several key contributions as listed below:

- Utilizes the Drug Bank dataset containing approximately 4,900 pharmaceutical compounds for predictive analysis of pharmaceutical pricing and composition.
- Improves data quality and model reliability through advanced preprocessing techniques, ensuring better learning performance from complex biomedical data.

- Demonstrates the effectiveness of ensemble learning techniques in handling complex pharmaceutical datasets.
- Develops an optimized AdaBoost ensemble model that improves classification performance by combining multiple weak learners.
- Provides a reliable AI-driven framework for enhancing pharmaceutical product pricing and composition prediction.

B. Justification and Novelty

The increasing complexity and cost burden in the pharmaceutical sector necessitates intelligent models for accurate drug pricing and composition prediction. Traditional methods often fail to handle nonlinear relationships and imbalanced biomedical datasets effectively. This study introduces a novel ensemble-based AdaBoost classifier integrated with advanced preprocessing techniques such as normalization, encoding, and data balancing. The proposed approach enhances learning efficiency and improves predictive performance compared to existing single-model methods. These words emphasize the significance and innovation of the proposed framework in pharmaceutical data analysis.

C. Organization of the Paper

The rest of the paper is organized as follows: Section II reviews related work, Section III describes the dataset and proposed methodology, Section IV presents the experimental results and analysis, and Section V concludes the study with future research directions.

II. LITERATURE REVIEW

A thorough literature search of studies on Pharmaceutical Pricing and Composition Prediction was conducted to aid in the formulation of the proposed research. Table I provides a summary of the recent research works undertaken, along with the models used, datasets employed, important research findings and challenges.

Omoora and Nagem (2026) investigates the application of Automated Machine Learning (AutoML) on historical sales data to forecast supply needs. By automating model selection and hyperparameter optimization, the AutoML pipeline identifies effective forecasting models with minimal human intervention. The study evaluates various algorithms for supply forecasting, including time-series forecasting methods and machine-learning regression models, which use historical sales patterns to predict future supply. The developed categorisation-aware AutoML stacked ensemble reduced sMAPE by 54% on daily series (from 68.14% to 31.54%) and by 31% on weekly series (from 39.17% to 27.15%) across eight drugs, with the largest single-drug gain for N05C (daily sMAPE from 161.92% to 23.00%). These findings highlight the value of combining data categorization with AutoML to enhance forecasting accuracy, optimize supply chain management, and support data-driven decision-making in pharmaceutical distribution[22].

Nag and Helal (2025) evaluates the performance of statistical (Holt-Winters) and machine learning-based (Prophet) models for forecasting monthly pharmaceutical demand in Kuwait, using three years of Diphtheria-Tetanus-Pertussis vaccine data. The Holt-Winters multiplicative model identified a significant upward trend (Sen's slope: 5,495 units/month) and seasonal patterns but yielded moderate accuracy (MAPE: 25.39%, MSD: 51.7M), struggling with outliers like an unusual demand spike in a specific month. Prophet demonstrated superior training performance (MAPE: 18.2%) by capturing Kuwait's unique summer holiday effects through its additive decomposition framework. However, its test accuracy declined (MAPE: 31.9%), reflecting challenges in generalizing to abrupt trend shifts and limited data. Both models highlighted the dominance of unmodeled factors (78% variability unexplained by trend alone), emphasizing the need for larger datasets and hybrid approaches. While Prophet excels in handling complex seasonality, its sensitivity to data quality and volume underscores practical limitations[23].

Malathi et al. (2024) proposed study highlights the need for a revolutionary strategy to overcome the limitations of existing systems. Quality control systems in existing systems have limitations such as real-time monitoring, insufficient data analysis for pattern identification, and limited predictive capabilities. These limitations have an impact on the industry's ability to recognize and manage quality concerns, posing a danger to medicine manufacturing. For addressing these issues, a novel approach to maintain quality control in the pharmaceutical sector using big data analytics is proposed. Advanced analytics are used for real-time monitoring, data analysis and predictive modeling in the proposed system. The results and analysis through extensive comparisons shows proposed big data analytics system achieves 95% fault detection, 40% downtime reduction, and 35% waste reduction in pharmaceutical manufacturing. The proposed system's ability to handle massive datasets allows it to identify complicated patterns and abnormalities, enabling proactive quality control. The implementation exceeds industry standards, addressing data privacy and security concerns[24].

Kumar (2024) proposed to incorporates various features like users' health condition, drug ratings and intake dates with the objective of interpreting drug users' sentiments through their feedback. carefully analyzed the data using method like descriptive statistics and data visualization followed by testing various machine learning model such as Light GBM model, Random Forest Classifier, and LSTM Model and found the accuracy ranging from 90% to 95.7%. used the dataset from the UCI machine learning repository with over 161297 reviews and 7 features. Since the reviews didn't have sentiment labels came up with the new method, assigned sentiment labels on the ratings given by the user either positive or negative. This helped us uncover the hidden sentiment in the reviews. Overall, this research study highlights how people feel about drugs and its impact on experience and treatment outcomes[25].

Lam et al. (2023) introduse the team embarked on utilizing these techniques to forecast the consumption of essential medications in Vietnam, particularly within the context of Nghe An province. employed machine learning models such as Random Forest, LGBM, Histogram-Based Gradient Boosting, and XGBoost, focusing on the Nghe An province. By leveraging specific pharmaceutical transaction data in the region, gained valuable insights into the dataset characteristics through data collection, preprocessing, and data analysis. aimed to predict the usage demands for commonly used drugs by formulating a regression problem. initial results showed great promise, with the highest Root Mean Square Error (RMSE) of 0.95 and both R-squared and Adjusted R-squared values of 0.81 belonging to Random Forest. Through continuous endeavors, aspire to enhance the accuracy and

effectiveness of drug demand forecasting in the pharmaceutical sector and then optimize inventory management, improve resource allocation, and enhance service delivery to meet the healthcare needs of the people of Nghe An province and Vietnam as a whole[26].

Dutta, Das and Chatterjee (2022) proposed to predict sales accurately, use different machine learning algorithms. can find complicated patterns in the sales dynamics including various risk variables in detailed study and analysed comprehensible predictive models to improve future sales predictions. Building a model based on historical data to forecasting sales of medicines, which can be applicable to new drugs which are licensed and released for sales. A way to show the effectiveness of the forecasting sales in drugs, taking the factors influencing, revealing the reviews of the existing solutions and analysing specific areas. have tested with 5 different machine learning algorithms with the pharmaceutical product dataset and reached to a best algorithm i.e. linear regression. Its performance, Mean absolute percentage error (MAPE) is 19.07% and is better than other performing model. Hence experiment shows the linear regression model is the best model for predicting pharmaceutical product sales[27].

Konar and Pitroda (2022) aim to built two prediction models that use supervised learning which would predict how likely a person utilize online pharmacies and at-home lab tests post COVID 19 pandemic. It consists of 5 probabilities i.e. very likely to use it, likely use it, moderately use it, less likely to use it, and rarely use it. The various classification algorithms employed in the study were logistic regression, decision trees, random forest, support vector machines, and Gradient booster classification algorithm The model's results are based on the previous data that was collected by asking around 250 individuals. Accuracies obtained for predicting how likely a person use online pharmacies post COVID are as follows, by gradient booster model the accuracy is ~85%, the accuracy obtained by decision tree classifier is ~71%, by logistic regression model the accuracy is ~85% and accuracy obtained by random forest model is ~78%. Accuracies obtained for predicting how likely a person use at-home lab tests post COVID are as follows, by decision tree classifier accuracy is ~78%, the accuracy obtained by logistic regression is ~50% and the accuracy obtained by support vector machines is ~64%[28]

TABLE I. RECENT STUDIES ON PHARMACEUTICAL PRICING AND COMPOSITION PREDICTION USING MACHINE LEARNING TECHNQIUES

Author	Proposed Work	Results	Key Findings	Limitations & Future Work
Omoora and Nagem (2026)	Developed a categorization-aware AutoML stacked ensemble for pharmaceutical supply forecasting using historical sales data.	Daily sMAPE reduced from 68.14% to 31.54% and weekly sMAPE from 39.17% to 27.15%.	AutoML combined with data categorization significantly improves pharmaceutical supply forecasting accuracy.	Limited to selected drugs and historical sales data; future work should incorporate external market and seasonal factors.
Nag and Helal (2025)	Compared Holt-Winters and Prophet models for monthly pharmaceutical demand forecasting.	Prophet achieved 18.2% training MAPE, while Holt-Winters obtained 25.39% MAPE.	Prophet effectively captures complex seasonal patterns and demand fluctuations.	Performance decreases with limited data and abrupt demand changes; hybrid forecasting approaches are recommended.
Malathi et al. (2024)	Proposed a big data analytics framework for pharmaceutical quality control and monitoring.	Achieved 95% fault detection, 40% downtime reduction, and 35% waste reduction.	Real-time analytics improves manufacturing quality and operational efficiency.	Requires large-scale infrastructure and high-quality data; future work can integrate AI-driven optimization techniques.
Kumar (2024)	Applied ML and LSTM models for pharmaceutical sentiment analysis using drug reviews.	Achieved classification accuracy between 90% and 95.7%.	Drug reviews provide valuable insights into patient experiences and treatment outcomes.	Sentiment labels were generated from ratings; future studies can employ advanced NLP and transformer-based models.
Lam et al. (2023)	Developed machine learning models for forecasting medicine demand in Vietnam.	Random Forest achieved RMSE = 0.95 and R^2 = 0.81.	Machine learning models can effectively predict pharmaceutical demand and support inventory management.	Dataset was limited to a specific region; future work should utilize larger and more diverse datasets.
Dutta, Das and Chatterjee (2022)	Evaluated multiple machine learning algorithms for pharmaceutical product sales forecasting.	Linear Regression achieved the best performance with MAPE = 19.07%.	Historical sales data can effectively support medicine sales prediction.	Limited consideration of external factors influencing sales; future research should incorporate economic and market variables.
Konar and Pitroda (2022)	Developed supervised learning models to predict post-COVID adoption of online pharmacies and at-home lab testing services.	Gradient Boosting and Logistic Regression achieved approximately 85% accuracy.	Machine learning can successfully model consumer healthcare behavior and technology adoption.	Small survey dataset limits generalizability; future work should include larger and more diverse populations.

Research gaps: Even though there has been considerable progress in pharmaceutical forecasting and analytics, there are still some gaps in the literature. The current research that exists is mostly related to demand forecasting, sales prediction or sentiment analysis, not directly pharmaceutical pricing and composition analysis. Much of the literature is based on restricting data sets or using a single modeling method, leading to a lower level of predictive robustness and generalization. Furthermore, few studies combine full preprocessing with ensemble learning methods to effectively deal with the complexity of data and balance. Thus, there is a need for a better AI-based framework that can provide better accuracy and reliability for pharmaceutical product prediction problems.

III. RESEARCH METHODOLOGY

The methodology proposed here employ the DrugBank data set, which was pre-processed using methods such as handling missing data, removing duplicate data, One-Hot Encoding, Z-Score normalization, and Random Over Sampler

for balancing classes. The processed data is then divided into 80% training data and 20% testing data, which is used to create and test the model. Then an AdaBoost ensemble classifier is employed and its performance is evaluated based on accuracy, precision, recall, F1 score and ROC curve analysis. The proposed flowchart of Pharmaceutical Pricing and Composition Prediction is shown in Fig. 1

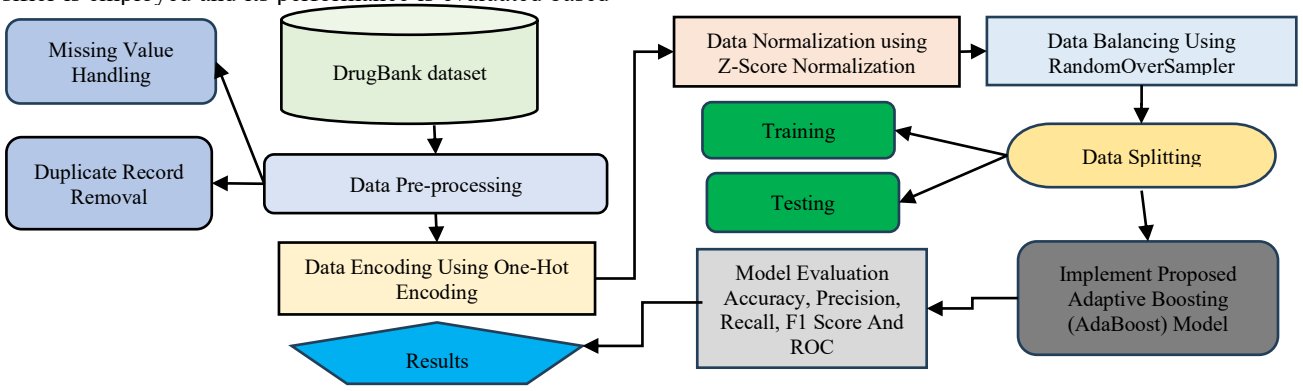

Fig. 1. Proposed Flowchart for Pharmaceutical Pricing and Composition Prediction using Machine Learning

The following section presents a detailed explanation of each step involved in the proposed methodology:

A. Data Gathering and Analysis

This study utilizes the Drug Bank dataset, a pharmaceutical repository, The dataset includes approximately 4,900 pharmaceutical compounds, including drug names, Drug Bank IDs, chemical compositions, and interaction details. Data visualization techniques such as bar plots and heatmaps were used to analyze data distribution and examine feature correlations:

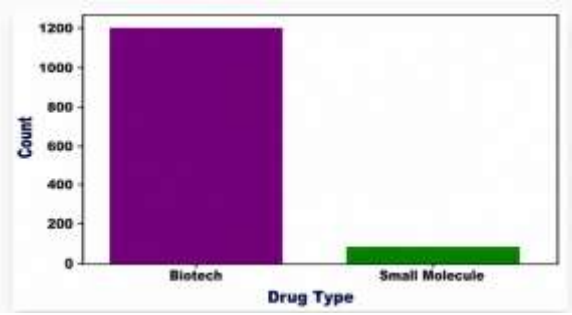

Fig. 2. Bar Graph of Class Distribution

Fig. 2 illustrates the distribution of drug types within the DrugBank dataset. As seen, the Biotech category has many more records than the small Molecule category, hence the class distribution is imbalanced. This visualization helps to grasp the composition of the dataset, and to prepare it appropriately for model training and to balance it if needed.

B. Data Pre-processing

The Drug Bank dataset was utilized for data preparation, which involved data integration, cleaning, and preprocessing. Key preprocessing steps included missing value handling, duplicate record removal, data labeling, and normalization to improve data quality and ensure suitability for machine learning analysis. The main preprocessing procedures are summarized as follows:

- **Missing Value Handling:** Rows containing missing values in the cucumber attribute were removed to maintain data consistency. Missing entries in the uniprot column were filled using synthetic or generated values to preserve dataset completeness.
- **Duplicate Record Removal:** Duplicate drug records were identified and eliminated to avoid data redundancy and reduce bias during model training. This improved the reliability of the dataset.

C. Data Encoding Using One-Hot Encoding

One-hot encoding converts categorical variables into binary vectors, where each category is represented by a separate column containing values of 0 or 1. This transformation enables machine learning algorithms to process categorical data effectively without introducing ordinal relationships between categories.

D. Data Normalization using Z-Score Normalization

Data Normalization is a technique used to transform or standardize the data to have a similar distribution. The most widely used method for data normalization is rescaling or min–max normalization and z-score normalization. In this study, have applied z-score normalization, a standardization technique with a mean of 0 and a standard deviation of 1. This scaling technique transforms values centered around the average, with a unit standard deviation. The z-score normalization is defined as given in Equation (1).

$$E' = \frac{E-\bar{M}}{\sigma_M} \quad (1)$$

Where,

E' and E are new and old for each data entry, $\bar{M}$ is the mean, and σ_M is the standard deviation.

E. Data Balancing Using RandomOverSampler

Data balancing is the process of addressing an imbalanced dataset, where one class has significantly fewer samples than the others, by adjusting the distribution of classes to improve the machine learning model's performance. RandomOverSampler is a method used for data balancing inmachine learning when dealing with imbalanced datasets. It is frequently applied to deal with the problem of significantlymore samples in one class than in another, which might result in biased model performance. The minority class, or the class with fewer instances, is the sample source used by RandomOverSampler to generate new synthetic samples by randomly duplicating samples from that class. Up until the minority class's sample count equals that of the majority class' sample count, this process is repeated. After implementing the

random oversample on the dataset, get the 958 values for class 1 and 936 for class 0.

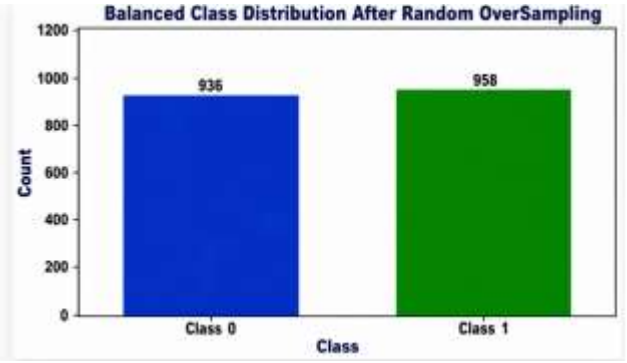

Fig. 3. Bar graph of class distribution after RandomOverSampling

Fig. 3 illustrates the class distribution after applying Random Oversampling, showing that both classes have nearly equal numbers of samples. This balanced distribution (Class 0: 936 and Class 1: 958) helps reduce class imbalance and improves the reliability of ML model training.

F. Data Splitting

To increase accuracy and efficacy for this phase, the undersampled data are divided into training and testing data, keeping a ratio of 80% training data and 20% testing data. After splitting, the model is trained using the different ensemble classifiers.

G. Proposed Adaptive Boosting (AdaBoost) Model

In this work, a supervised ML-based boosting model, namely the Proposed Adaptive Boosting (AdaBoost) Model, is developed for pharmaceutical pricing and composition prediction. AdaBoost (Adaptive Boosting) is an ensemble learning algorithm that improves classification performance by combining multiple weak learners into a single strong classifier. The model trains weak classifiers sequentially, with each classifier focusing on samples incorrectly classified by the previous learners. This iterative process enables AdaBoost to minimize classification errors and enhance predictive accuracy. The final strong classifier is formed by aggregating the weighted outputs of all weak learners, as expressed in Equation (2).

$$F(x) = \sum_{t=1}^{T} \alpha_t h_t(x) \quad (2)$$

where $F(x)$ denotes the final strong classifier, $h_t(x)$ represents the weak learner at iteration t, α_t is the weight assigned to the corresponding weak learner, and T is the total number of boosting iterations.

The effectiveness of AdaBoost depends on the contribution of each weak classifier. Therefore, the algorithm calculates a weight for every classifier based on its classification error. Weak learners with lower error rates receive higher weights, allowing them to contribute more significantly to the final decision. The weight of the t^{th} weak classifier is computed using Equation (3).

$$\alpha_t = \frac{1}{2} In\left(\frac{1-\epsilon_t}{\epsilon_t}\right) \quad (3)$$

Where α_t is the importance weight of the t^{th} weak classifier and ϵ_t denotes its weighted classification error. These classifier weights are then utilized to construct the final ensemble model, improving the overall classification performance and robustness of the AdaBoost algorithm.

The proposed AdaBoost model uses Decision Tree as the base estimator with a maximum depth of 3 and 100 estimators in the ensemble. The learning rate is set to 1.0 to balance the contribution of each weak learner during training. The model is trained using the SAMME.R algorithm for improved probabilistic boosting performance. These hyperparameter values are selected to ensure optimal classification accuracy and model stability on the pharmaceutical dataset.

H. Evaluation Metrics

Various classification parameters were used to assess the performance of the proposed model. A confusion matrix was first created to investigate the prediction results based on the comparison between prediction class labels and real class labels. The values of True positives (TP), False positives (FP), True negatives (TN) and False negatives (FN) were provided in the matrix and were used to calculate some important evaluation metrics like accuracy, precision, recall and F1-score. These measures give us a full picture of the accuracy of the model's classification and its reliability in prediction:

Accuracy: Accuracy represents the proportion of correctly classified instances to the total number of instances in the dataset. It measures the overall effectiveness of the model in making correct predictions and is calculated using Equation (4)-

$$Accuracy = \frac{\text{TP+TN}}{\text{TP+Fp+TN+FN}} \quad (4)$$

Precision: Precision is the ratio of correctly predicted positive instances to the total number of instances predicted as positive by the model. It measures the classifier's ability to accurately identify positive cases while minimizing false positive predictions and is expressed in Equation (5)-

$$Precision = \frac{\text{TP}}{\text{TP+FP}} \quad (5)$$

Recall: Recall is the ratio of correctly predicted positive instances to the total number of actual positive instances in the dataset. It measures the model's ability to identify all relevant positive cases and is mathematically expressed in Equation (6)-

$$Recall = \frac{\text{TP}}{TP+FN} \quad (6)$$

F1 score: The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's classification performance. It is particularly useful when both false positives and false negatives are important, and its value ranges from 0 to 1. The mathematical expression for the F1-score is given in Equation (7)-

+ln

Receiver Operating Characteristic Curve (ROC): The Receiver Operating Characteristic (ROC) curve is a graphical tool used to evaluate the classification performance of a model across different decision thresholds. It plots the True Positive Rate (TPR) against the False Positive Rate (FPR), where TPR is also known as sensitivity or recall, and FPR is calculated as 1 − specificity. A ROC curve closer to the upper-left corner indicates better classification performance.

IV. RESULTS AND DISCUSSION

The experimental setup and performance evaluation of the proposed AdaBoost model are provided in this section, both during training and testing for pharmaceutical pricing and composition prediction. Experiments have been performed on

a Windows 10 (version 20H2) computer with an Intel i7-11800H CPU (4.60 GHz) and 64 GB RAM. All implementations were written in Python 3.9.12 and chemical compound representations were created with PyBioMed (PyInteraction module) and RDKit (version 1.0.3). The proposed model was trained and tested for the DrugBank database with some standard performance metrics such as accuracy, precision, recall and F1 score. The AdaBoost model showed 99% accuracy, precision, recall, and F1-score as shown in Table II, which indicates that it exhibited an excellent classification ability, high reliability and good predictive performance for the prediction of pharmaceutical pricing and composition.

TABLE II. CLASSIFICTION RESULTS OF PROPOSED ADAPTIVE BOOSTING MODEL PHARMACEUTICAL PRICING AND COMPOSITION PREDICTION

Matrix	Testing	Training
Accuracy	99	100
Precision	99	100
Recall	99	100
F1-score	99	100

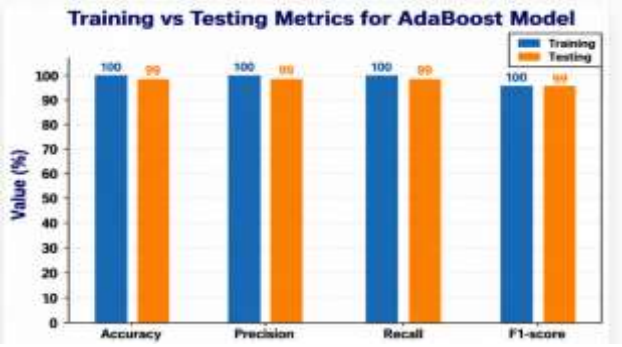

Fig. 4. Train and Test Performance Comparison of Proposed Model

Fig. 4 illustrates the bar chart compares the training and testing performance of the proposed AdaBoost model using Accuracy, Precision, Recall, and F1-score metrics. The model achieved 100% performance on the training dataset and 99% performance on the testing dataset with each of the metrics, showing excellent prediction capability, good generalization and minimal overfitting in the pharmaceutical pricing and composition prediction.

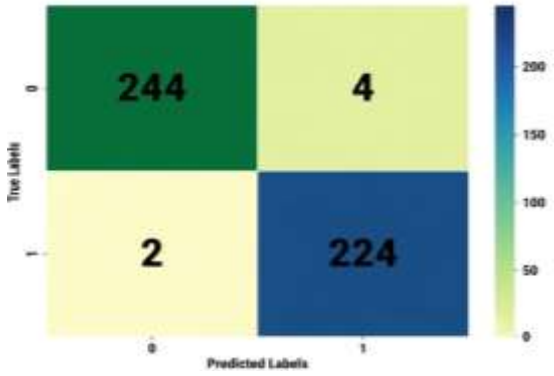

Fig. 5. Confusion Matrix for the Proposed AdaBoost Model

Fig. 5 presents This diagram represents the confusion matrix for the proposed AdaBoost binary classification model, mapping actual true labels against the model's predicted labels for classes 0 and 1. The visualization uses a new high contrast gradient ranging from a soft light yellow for low counts to rich greens and deep blues for peak counts, highlighting a very accurate diagonal performance distribution. The model has an exceptionally high true negative rate and true positive rate; 244 instances of class 0 were correctly classified and 224 instances of class 1 were correctly classified; and it had minimal classification errors with 4 false positive and 2 false negatives, total.

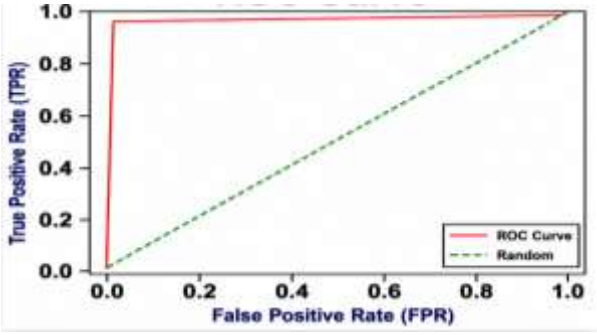

Fig. 6. ROC Curve for Proposed AdaBoost Model

Fig. 6. ROC Curve for Proposed AdaBoost Model: This shows the classification performance of the proposed AdaBoost model with the True Positive Rate (TPR) and False Positive Rate (FPR). The curve remains close to the upper-left corner and significantly above the random classification line, indicating excellent discriminative ability and a high Area Under the Curve (AUC). This shows that the AdaBoost model discriminates between the classes and can have excellent predictive accuracy in the prediction of the price and composition of the pharmaceutical products.

A. Comparative Analysis

To assess the effectiveness of the proposed AdaBoost model, a comparative evaluation was conducted against several existing machine learning models, including KNN, SVM, CNN, and XGBoost. Table III gives a performance comparison; it can be seen that the proposed AdaBoost classifier obtained the best performance in all the performance measures. In detail, it achieved 99% accuracy, precision, recall and F1-score, which is better than XGBoost (97.1%) and CNN (94.6%). The progressive improvement in all metrics shows that the model can robustly detect the patterns of pharmaceutical composition and pricing-related attributes. The results show that the proposed AdaBoost strategy is quite robust, reliable, and has great predictive power in the DrugBank dataset.

TABLE III. COMPARISON OF DIFFERENT MACHINE LEARNING MODELS FOR PHARMACEUTICAL PRICING AND COMPOSITION PREDICTION

Model	Accuracy	Precision	Recall	F1-score
KNN[29]	74.05	72.5	70.7	74.8
SVM[30]	78.5	82	62	78
CNN[31]	94.6	92.06	97.7	94.8
XGBoost[8]	97.1	98.2	87	92.3
Proposed AdaBoost Classifier	99	99	99	99

The proposed AdaBoost classifier obtained an excellent accuracy of 99%, showing its effectiveness to predict the pharmaceutical pricing and composition with high reliability and accuracy. The high precision, recall, and F1-score values of the classification model suggest that it has made only a few prediction errors and achieved good classification accuracy. In addition, the ensemble learning mechanism of AdaBoost makes the model more robust by stacking multiple weak learners, thereby better generalizing and better dealing with

complex data patterns and more effectively predicting compared to traditional machine learning methods.

V. Conclusion and Future Study

Improving pharmaceutical product pricing and predicting their composition is a process that can be done accurately using Artificial Intelligence-based analysis, which helps in data-driven decisions. It improves the capacity to analyze intricate drug data sets and find patterns that can help make more accurate predictions. The machine learning models evaluated in this study included different models to determine which would be the most effective for classification tasks. The results for the model indicate that the proposed AdaBoost classifier outperforms the other existing classifiers like KNN (74.05%), SVM (78.5%), CNN (94.6%) and XGBoost (97.1%) with highest accuracy of 99%. It shows that ensemble-based AdaBoost model has excelled in capturing complex relationships in pharmaceutical data and in providing highly reliable predictive performance.

The study was conducted on a single dataset, so its results might not be applicable in other pharmaceutical datasets. It also can encounter real world data on which it is very hard to obtain clear meaning. Larger multi-source datasets can be used in future to increase the robustness. Furthermore, hybrid techniques involving deep learning and explainable AI techniques can be investigated to achieve enhanced performance and interpretability.

References

[1] H. Zhao, X. Yao, Z. Liu, and Q. Yang, "Impact of Pricing and Product Information on Consumer Buying Behavior With Customer Satisfaction in a Mediating Role," *Front. Psychol.*, vol. 12, Dec. 2021, doi: 10.3389/fpsyg.2021.720151.

[2] R. S. Snehamruth, "Data-Driven Optimization of Pharmaceutical Manufacturing Processes using Quality by Design (QbD) Frameworks," *Int. J. Curr. Eng. Technol.*, vol. 14, no. 6, pp. 557–566, 2024, doi: 10.14741/ijcet/v.14.6.19.

[3] B. A. Mousa and B. Al-Khateeb, "Predicting medicine demand using deep learning techniques: A review," *J. Intell. Syst.*, vol. 32, no. 1, p. 20220297, 2023, doi: doi:10.1515/jisys-2022-0297.

[4] A. Tomovic and E. Atukeren, "Long-term value creation in the pharmaceutical sector: an event study analysis of big pharma stocks," *Int. J. Sustain. Econ.*, vol. 4, no. 4, p. 370, 2012, doi: 10.1504/IJSE.2012.049609.

[5] P. Kumar, "Leveraging Generative AI for Automated Data Standardization and Interoperability in Healthcare," in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Dec. 2025, pp. 99–104. doi: 10.1109/ICAAIC64647.2025.11330217.

[6] M. Fazekas, Z. Veljanov, and A. B. de Oliveira, "Predicting pharmaceutical prices. Advances based on purchase-level data and machine learning," *BMC Public Health*, vol. 24, no. 1, p. 1888, Jul. 2024, doi: 10.1186/s12889-024-19171-9.

[7] E. D. Kantor, C. D. Rehm, J. S. Haas, A. T. Chan, and E. L. Giovannucci, "Trends in Prescription Drug Use Among Adults in the United States From 1999-2012," *JAMA*, vol. 314, no. 17, p. 1818, Nov. 2015, doi: 10.1001/jama.2015.13766.

[8] Q.-H. Kha, V.-H. Le, T. N. K. Hung, N. T. K. Nguyen, and N. Q. K. Le, "Development and Validation of an Explainable Machine Learning-Based Prediction Model for Drug–Food Interactions from Chemical Structures," *Sensors*, vol. 23, no. 8, p. 3962, Apr. 2023, doi: 10.3390/s23083962.

[9] H. N. Dholariya, "AI-Governed Data Modernization Architectures: A Secure and Compliant Framework for Healthcare and Life Sciences Cloud Ecosystems," *Front. Heal. Informatics*, vol. 15, no. 1, pp. 102–117, Apr. 2026, doi: 10.63682/fhi2984.

[10] R. K. Gadiraju, "AI-Driven Hardware Telemetry Architecture for Predictive Device Health," *Int. J. Recent Innov. Trends Comput. Commun.*, vol. 11, no. 11, pp. 2049–2056, Nov. 2023, doi: 10.17762/ijritcc.v11i11.11947.

[11] A. Warrier, "Real-Time AI Integration Architectures for HIPAA-Compliant Healthcare Data Interoperability," in *International Journal of Emerging Trends in Computer Science and Information Technology*, Eureka Vision Publication, 2025, pp. 74–81. doi: 10.63282/3050-9246/WCAI25-128.

[12] N. Kavitha and P. Madhumathy, "Real-time pill identification and classification using deep learning framework for medicine inspection systems," *Discov. Electron.*, vol. 2, no. 1, p. 80, Oct. 2025, doi: 10.1007/s44291-025-00122-6.

[13] B. Munos, "Lessons from 60 years of pharmaceutical innovation," *Nat. Rev. Drug Discov.*, vol. 8, no. 12, pp. 959–968, Dec. 2009, doi: 10.1038/nrd2961.

[14] K.-K. Mak and M. R. Pichika, "Artificial intelligence in drug development: present status and future prospects," *Drug Discov. Today*, vol. 24, no. 3, pp. 773–780, Mar. 2019, doi: 10.1016/j.drudis.2018.11.014.

[15] R. Sharma, A. Shishodia, A. Gunasekaran, H. Min, and Z. H. Munim, "The role of artificial intelligence in supply chain management: mapping the territory," *Int. J. Prod. Res.*, vol. 60, no. 24, pp. 7527–7550, Dec. 2022, doi: 10.1080/00207543.2022.2029611.

[16] IJSRCSEIT and Dr. Harikrishna Jethva, "International Journal of Scientific Research in Computer Science, Engineering and Information Technology," *Int. J. Sci. Res. Comput. Sci. Eng. Inf. Technol.*, vol. 1, no. 1, p. 1, Jan. 2017, doi: 10.32628/IJSRCSEIT.

[17] P. Mishra, G. G. Ramani, M. I. Patel, M. S. Soumik, S. Mahmud, and R. Manivannan, "Design of intelligent healthcare IT infrastructure using graph theory, network analysis, and artificial intelligence," *Int. J. Appl. Math.*, vol. 38, no. 12s, pp. 2267–2280, 2025.

[18] N. R. C. Monteiro, B. Ribeiro, and J. P. Arrais, "Drug-Target Interaction Prediction: End-to-End Deep Learning Approach," *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 18, no. 6, pp. 2364–2374, Nov. 2021, doi: 10.1109/TCBB.2020.2977335.

[19] L. Xu, X. Ru, and R. Song, "Application of Machine Learning for Drug–Target Interaction Prediction," *Front. Genet.*, vol. 12, Jun. 2021, doi: 10.3389/fgene.2021.680117.

[20] H. Hanafiah, "Dampak Financial Technology (Fintech) Terhadap Perkembangan Produk Bank Syariah Di Kota Bukit Tinggi," *Front. Neurosci.*, 2021.

[21] P. Kumar, "Edge Computing and IoT for Real-Time Healthcare Data Processing and Integration," in *2025 4th International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, IEEE, Dec. 2025, pp. 105–110. doi: 10.1109/ICAAIC64647.2025.11331211.

[22] E. S. Omoora and T. Nagem, "Supply Forecasting for a Pharmaceutical Distribution Company Using AutoML Techniques on Historical Sales Data," in *2026 IEEE 5th International Maghreb Meeting of the Conference on Sciences and Techniques of Automatic Control and Computer Engineering (MI-STA)*, IEEE, Apr. 2026, pp. 492–497. doi: 10.1109/MI-STA68962.2026.11511212.

[23] K. Nag and M. Helal, "Monthly Pharmaceutical Demand Forecasting Using Statistical Holt Winters Model and Machine Learning-based Prophet Model," in *2025 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM)*, IEEE, Dec. 2025, pp. 0539–0543. doi: 10.1109/IEEM63636.2025.11357790.

[24] P. Malathi, S. Jansi, M. V. Prabhakaran, K. Senbagam, M. P. Sujatha, and R. Sadaieswaran, "Big Data Analytics for Quality Control in The Pharmaceutical Industry," in *2024 10th International Conference on Advanced Computing and Communication Systems (ICACCS)*, IEEE, Mar. 2024, pp. 669–674. doi: 10.1109/ICACCS60874.2024.10717314.

[25] A. Kumar, "Machine Learning Based Sentiment Analysis of Pharmaceutical Reviews," in *2024 5th International Conference on Electronics and Sustainable Communication Systems (ICESC)*, IEEE, Aug. 2024, pp. 1850–1855. doi: 10.1109/ICESC60852.2024.10689892.

[26] L. D. Lam, B. P. Le Luong, H. T. Mai Linh, and P. M. Hung, "Application of Machine Learning in Predicting the Amount of Pharmaceutical Drugs Ordered for the Manufacturer," in *2023 1st International Conference on Health Science and Technology*

(ICHST), IEEE, Dec. 2023, pp. 1–6. doi: 10.1109/ICHST59286.2023.10565367.

[27] S. R. Dutta, S. Das, and P. Chatterjee, "Smart Sales Prediction of Pharmaceutical Products," in *2022 8th International Conference on Smart Structures and Systems (ICSSS)*, IEEE, Apr. 2022, pp. 1–6. doi: 10.1109/ICSSS54381.2022.9782271.

[28] K. Konar and H. Pitroda, "Analyzing and Predicting the Impact of COVID-19 on Online Pharmaceuticals Sectors and Pathological Services in India," in *2022 IEEE 7th International conference for Convergence in Technology (I2CT)*, IEEE, Apr. 2022, pp. 1–7. doi: 10.1109/I2CT54291.2022.9824883.

[29] X. Chen, Z. Wang, Z. Miao, and B. Nie, "Research on drug-drug interaction prediction using capsule neural network based on self-attention mechanism," *BMC Bioinformatics*, vol. 26, no. 1, p. 293, Dec. 2025, doi: 10.1186/s12859-025-06308-9.

[30] H. Ibrahim, A. M. El Kerdawy, A. Abdo, and A. Sharaf Eldin, "Similarity-based machine learning framework for predicting safety signals of adverse drug–drug interactions," *Informatics Med. Unlocked*, vol. 26, p. 100699, 2021, doi: 10.1016/j.imu.2021.100699.

[31] J. Zhu, C. Che, H. Jiang, J. Xu, J. Yin, and Z. Zhong, "SSF-DDI: a deep learning method utilizing drug sequence and substructure features for drug–drug interaction prediction," *BMC Bioinformatics*, vol. 25, no. 1, p. 39, Jan. 2024, doi: 10.1186/s12859-024-05654-4.

