Pharmaceutical Sales Forecasting using Machine Learning

Jacob George
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering (Autonomous)
Kanjirappally, Kottayam, India
jacobgeorge2026@cs.ajce.in

Jerin Xavier
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering (Autonomous)
Kanjirappally, Kottayam, India
jerinxavier2026@cs.ajce.in

Jovin J George
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering (Autonomous)
Kanjirappally, Kottayam, India
jovinjgeorge2026@cs.ajce.in

Joyel Xavier
Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering (Autonomous)
Kanjirappally, Kottayam, India
joyelxavier2026@cs.ajce.in

Subini Therese Babu
Faculty, Dept. of Computer Science & Engineering
Amal Jyothi College of Engineering (Autonomous)
Kanjirappally, Kottayam, India
subinitheresebabu@amaljyothi.ac.in

***Abstract*—Accurate pharmaceutical sales forecasting is crucial for managing inventory, improving supply chains, and lowering financial risks from stockouts and product expiries. Traditional statistical methods like ARIMA often struggle to address nonlinear dependencies and irregular demand patterns found in real retail environments. In response, recent research has increasingly used machine learning approaches that show better accuracy and flexibility. This paper reviews a collection of recent studies on time series forecasting, focusing on methods for data preprocessing, feature engineering, and model development. Based on the findings from these studies, the paper presents a structured forecasting perspective that combines effective preprocessing strategies with machine learning techniques suited for diverse pharmaceutical datasets. Special emphasis is placed on tree-based ensemble models like XGBoost for managing structured retail data and neural network methods for situations with limited historical records. The discussion highlights how these complementary techniques can work together to tackle challenges such as demand fluctuations, sparse data conditions, and support for operational decisions in pharmaceutical supply chains. Comparative results from the studies underscore the reliability of XGBoost in handling structured datasets and the performance of GRNN in low-data scenarios. The discussion also addresses key limitations, such as interpretability and scalability, and suggests future directions for real-world application. Overall, the study shows that machine learning models, particularly ensemble and neural network approaches, offer a promising route to reliable and actionable pharmaceutical sales forecasting.**

***Index Terms*—Time series forecasting, pharmaceutical sales, XGBoost, GRNN, machine learning, data preprocessing, retail analytics, cold-start forecasting.**

I. Introduction

Sales forecasting is very important in the pharmaceutical sector. Demand changes, seasonal variations, and product life cycles all affect supply chain efficiency and patient access. Unlike sales predictions in general retail, forecasting for pharmaceuticals is more complex. New drug launches can face cold-start issues, there are regulatory requirements, and sales data is often fragmented across distributors, pharmacies, and regions. In this context, dependable forecasting systems are crucial for inventory planning, managing expirations, and helping stakeholders make informed decisions. [1].

Traditional statistical forecasting methods, like ARIMA and SARIMA, are commonly used in time series analysis. They work well for stationary and organized datasets, but often face challenges with noisy, sparse, and irregular retail data. Recent progress in machine learning (ML) and deep learning (DL) has led to more flexible models that can capture nonlinear

patterns and seasonality. Gradient boosting methods, such as XGBoost and LightGBM, have shown better performance in structured retail datasets [5], while researchers have looked into architectures like LSTM and GRU for modeling sequences [13].However, research shows that deep learning models may not work as well with fragmented, small-scale datasets. This is due to their high data needs and their tendency to overfit. [2].

Another pressing challenge in pharmaceutical forecasting is the *cold-start problem*. New products do not have enough historical data for conventional models. In these situations, lightweight neural architectures like General Regression Neural Networks (GRNN) and multilayer perceptrons (MLP) perform better than both statistical baselines and more complex deep networks. They provide a good balance between accuracy and efficiency in computation. [15], [17].

Preprocessing and feature engineering are both essential. Inconsistent and incomplete datasets can hurt forecasting performance. Automated pipelines for managing missing values, outliers, and categorical encodings have been suggested to ensure consistency and reproducibility in forecasting studies [9], [10]. Additionally, recent reviews highlight that hybrid approaches, which combine statistical, machine learning, and deep learning techniques, are becoming effective solutions for the complexities of sales forecasting [4], [11].

With an emphasis on pharmaceutical sales, this paper offers a thorough analysis of current developments in machine learning for time series forecasting. We examine ensemble approaches, cold-start solutions, and data preprocessing strategies, emphasising their advantages and disadvantages in practical applications. Building on these insights, the paper presents best practices in data preparation, feature engineering, model selection, and practical integration within pharmaceutical decision-support systems by synthesising current methodological approaches into an organised analytical perspective. The study critically investigates how various modelling approaches can be methodically aligned to address real-world forecasting constraints in pharmacy and distributor networks rather than suggesting a single implementation.

II. Related Work

In both academic and professional settings, sales forecasting has been thoroughly researched using a variety of methods, from traditional statistical models to cutting-edge machine learning techniques. Because of their interpretability and solid mathematical underpinnings, traditional models like ARIMA and SARIMA have long been used in demand forecasting. Nevertheless, performance in highly volatile or fragmented datasets is limited by their dependence on linear assumptions [4]. On the other hand, machine learning techniques are becoming more and more popular because of their capacity to integrate heterogeneous features, capture nonlinear dependencies, and adjust to changing demand conditions. According to recent reviews, accurate models are necessary for accurate forecasting, but preprocessing, feature design, and hybridisation techniques also need to be carefully considered. [1].

A. Data Preprocessing

Robust forecasting requires efficient preprocessing of time-series data. Typical methods for enhancing model stability include feature scaling, normalization, imputation of missing values, and duplicate removal. Research has demonstrated that poor management of these problems can seriously impair model performance, frequently more so than algorithm selection itself [9]. In order to standardize workflows and minimize human intervention, automated preprocessing frameworks have emerged, allowing for faster deployment across multiple datasets [10].

Imputation techniques need to be adapted to the type of missingness, according to recent research. Time-series contexts frequently employ techniques like Expectation Maximization (EM), Multiple Imputation by Chained Equations (MICE), k-Nearest Neighbors (kNN), and spline interpolation, each of which offers trade-offs between computational complexity and accuracy. Traditional techniques like Z-score and Interquartile Range (IQR) are still frequently used for outlier detection, but Isolation Forest offers a more reliable substitute in high-dimensional datasets. Similarly, to stabilize model training and convergence, normalization methods like Box-Cox transformation, Z-score normalization, and Min-Max scaling are commonly used [9], [12].

The efficiency of preprocessing is further increased by feature extraction and selection. While more sophisticated techniques such as Neighbourhood Component Analysis (NCA) and Laplacian scores, enable the identification of the most informative variables, Principal Component Analysis (PCA) is frequently used to reduce dimensionality and noise. These procedures aid in streamlining the input space without compromising prediction performance, which is especially important for gradient boosting models such as XGBoost [20].

Preprocessing tasks are increasingly arranged as Directed Acyclic Graph (DAG) pipelines to increase modularity and reproducibility. By enforcing standardised task orderings (such as imputation $\rightarrow$ outlier detection $\rightarrow$ normalisation $\rightarrow$ feature selection), this structure facilitates controlled experimentation and makes cross-dataset benchmarking simpler [10]. Furthermore, real-time data cleaning at the source lowers latency and guarantees efficiency in distributed environments, which is why edge preprocessing has recently been introduced in IoT and healthcare contexts. These developments point to promising paths for pharmaceutical sales forecasting systems, which frequently depend on inventory and sales data that must be updated quickly.

In the domain of pharmaceuticals, the preprocessing needs are especially challenging, especially because of the hierarchical nature of the data, which is recorded at several levels (distributors, pharmacies, and types of products). Fourkiotis and Tsadiras [11] show the effectiveness of preprocessing in improving the predictive accuracy, especially in the domain of pharmaceutical sales. Irregularities, such as the presence of negative sales figures, indicating return or cancellation of sales, are also present.

Overall, this body of literature also serves to reinforce the fact that the quality of preprocessing is at least as important as model selection in terms of forecast accuracy, as asserted in [19]. With the ability to incorporate more advanced, automated, domain-specific, and modularized pipelines, such modern-day systems can become more robust, reproducible, and ready-to-deploy in the context of pharmaceutical supply chain management.

Table I presents an analytical comparison of forecasting paradigms identified in the literature, highlighting their strengths, limitations, and applicability to pharmaceutical sales forecasting.

B. Ensemble Methods with XGBoost

Among all the models of machine learning, tree-based ensemble models like Extreme Gradient Boosting (XGBoost) have become prominent in structured sales data sets. The capacity of XGBoost to handle non-linear relationships and diverse features makes it highly efficient for forecasting in the retail and pharmaceutical industries. For instance, large-scale applications like forecasting sales at Walmart have proven that XGBoost is superior to statistical models like ARIMA and SARIMA and even some deep learning models in a fragmented scenario [5], [6], [7].

One of the major benefits that can be derived from the use of the XGBoost algorithm is its ability to scale with efficiency in handling high-dimensional feature space that results from feature engineering. In addition to that, there have been studies conducted by Andrabi et al. [8], which have pointed out its ability to be used with engineered features for handling seasonality. All these benefits make a strong case for the selection of the algorithm in the present framework.

C. Cold-Start Forecasting

However, one of the issues that plagues pharmaceutical sales forecasting is the cold start problem, wherein a newly introduced drug does not have pre-existing market history. The traditional statistical approach and even some of the advanced approaches of machine learning are found wanting in such situations because they are based on pre-existing patterns and history. This is where the General Regression Neural Network (GRNN) comes in as a potential alternative because of its generalization capacity even with limited history available.

It is found that GRNN and similar models perform better than ARIMA and even long short-term memory (LSTM) when applied in situations with limited history available [17]. Dudek [17] also shows that GRNN is able to perform with similar accuracy and even better in terms of reduced training time, which is critical in situations of real-time decision-making in pharmacy operations. This is especially relevant in situations wherein pharmacy operations are forced to deal with newly introduced drugs and are required to forecast demand for them under such uncertainty.

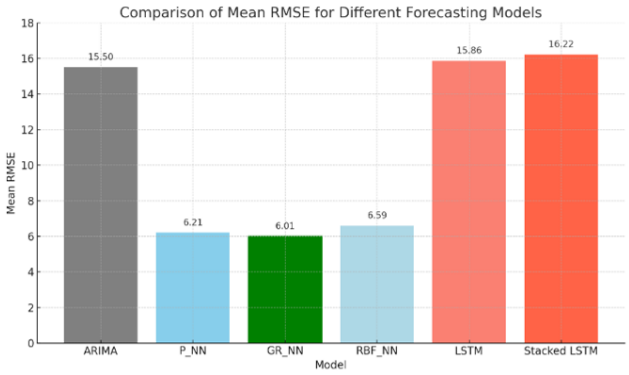

Fig. 1. Comparison of Mean RMSE values across forecasting models. The General Regression Neural Network (GRNN) achieves the lowest error, outperforming ARIMA and deep learning models such as LSTM and Stacked LSTM.

D. Comparative and Hybrid Approaches

Comparative studies have consistently revealed that no single model performs better than the others under all conditions. For instance, the LSTM and GRU models have been found to perform well for deep learning models with sequential data. However, the need for a larger training set and the computational complexity make them less suitable for the fragmented retail and pharmaceutical environments [2]. Conversely, gradient boosting models such as XGBoost have been found to be robust with moderate data sets and tuning complexity.

Hybrid models, which incorporate both statistical and machine learning models, have also started gaining more popularity. For instance, Bojer [4] suggests the decomposition of the time series into trend, seasonal, and residual components before applying the machine learning models. Such models have been found to be more viable as they incorporate the merits of both models, i.e., the interpretability of the statistical models and the flexibility of the machine learning models.

Yet, the literature does not commonly consider the following domain-specific requirements: inventory management with expo- nature of items, regulatory requirements, and incorporation into workflow within the pharmacy setting. Additionally, little consider- ation has been made for the following

TABLE I
ANALYTICAL COMPARISON OF FORECASTING PARADIGMS IN PHARMACEUTICAL SALES

Approach	Understanding from Existing Studies	Practical Strength	Common Limitations Reported
ARIMA / SARIMA (Statistical Models)	Treat forecasting as a linear time-series problem assuming stationarity and structured seasonality [17], [18].	Interpretable and mathematically grounded; effective for stable datasets.	Struggle with nonlinear patterns, abrupt demand shifts, and fragmented retail data [2], [4].
XGBoost (Ensemble Learning)	Models nonlinear interactions using gradient boosting over engineered temporal and categorical features [5]–[8].	Strong performance on structured retail datasets; scalable and efficient.	Requires careful feature engineering; reduced interpretability compared to statistical models [2].
LSTM / GRU (Deep Learning)	Capture long-term temporal dependencies using recurrent neural architectures [2], [13].	Effective for large sequential datasets with strong temporal patterns.	High data requirement; risk of overfitting in small or fragmented datasets [2].
GRNN (Cold-Start Neural Model)	Lightweight neural architecture capable of generalizing from limited samples [17].	Performs well in sparse-data environments; fast training.	Sensitive to smoothing parameter selection; limited long-sequence modeling ability [17].
Hybrid Approaches	Combine statistical decomposition with machine learning models to balance interpretability and nonlinear modeling [4].	Balanced performance and adaptability across forecasting conditions.	Increased pipeline complexity and integration challenges [4], [11].

deployment-related as- pects: scalability, interpretability, and real-time decision sup- port. It is evident that the need for forecasting strategies that not only have high predictive capability but also consider the practical and logistical requirements of pharmaceutical supply chain management has not been adequately addressed.

Table II summarizes the performance and deployment trade-offs across forecasting strategies, emphasizing practical considerations for real-world pharmaceutical applications.

III. DISCUSSION

The literature reviewed in this paper indicates the trend and reliance on machine learning models for time series forecasting in different domains. Statistical models such as ARIMA, although widely used in the past, are not able to capture nonlinear dependencies and irregular seasonality in complex retail and pharmaceutical data sets. On the other hand, ensemble models such as XGBoost have gained prominence as reliable alternatives, especially in capturing high-dimensional features and noisy data, and have been found to perform significantly better than linear and autoregressive models in different domains [5], [6]. This is also supported by the results of comparative studies, in which XGBoost has been found to perform better than deep learning models such as LSTM in structured sales environments characterised by engineered features and moderate data sizes [2], [5], [6].

In parallel to this, neural networks such as the General Regression Neural Network (GRNN) offer distinctive advantages in cold-start situations, as there is less availability of sales data. While recurrent networks require long-term dependencies to work effectively, the GRNN shows robust performance in data-constrained situations and can achieve lower RMSE compared to other models such as ARIMA and LSTM (Fig. 1) [13], [17]. This makes the model particularly useful in forecasting newly introduced pharmaceutical products, as rapid predictions are necessary in this scenario.

The other important aspect that has been emphasized by the various studies is the importance of the preprocessing step in the overall forecasting process. It has been highlighted that the handling of missing values, encoding, and the automation process play a vital role in improving the accuracy of the forecasting models [9], [10]. The implementation of advanced techniques such as kNN, MICE, Expectation Maximization, Isolation Forest, PCA, and NCA has further demonstrated that the accuracy of the forecasting models is equally dependent on the preprocessing step, with the quality of the data playing a vital role in the overall forecasting process [12], [20]. The importance of the reproducibility aspect has led to the development of the modular DAG pipeline, with edge preprocessing emerging as a vital aspect that facilitates real-time cleaning in the context of IoT and healthcare applications.

Literatures indicate that there is a potential balanced trade-off in terms of accuracy, robustness, and efficiency when ensemble learning techniques are integrated into structured data sets, while lightweight neural networks are integrated into sparse data sets. However, there is an underexplored but essential aspect of translating prediction outputs into decision-support systems, especially in the context of pharmaceutical supply chains, where inventory expiration and demand variability have a direct effect on the overall performance of the system. Thus, future research should not only focus on prediction but also on interpretability, scalability, and integration into real-world infrastructures.

IV. CONCLUSION

This review aims to synthesize information from sixteen recent publications on time series forecasting and present it from

TABLE II
PERFORMANCE AND DEPLOYMENT TRADE-OFF ACROSS FORECASTING STRATEGIES

Approach	Forecast Reliability	Data Requirement	Real-Time Feasibility	Scalability
ARIMA / SARIMA	Reliable for stable seasonal trends [17], [18].	Requires consistent historical time-series data.	Easy to implement in low-resource environments.	Scales moderately but struggles with high-dimensional features.
XGBoost	High predictive robustness in structured sales datasets [5], [6].	Moderate data volume with engineered features.	Suitable for near real-time inference with optimized pipelines.	Highly scalable due to parallelized boosting architecture.
LSTM / GRU	Strong in capturing long sequential dependencies [13].	Large labeled datasets required.	May require hardware acceleration for strict time constraints.	Scaling increases computational and infrastructure demands.
GRNN	Stable performance in cold-start and sparse scenarios [17].	Very limited historical data needed.	Fast training; suitable for rapid deployment.	Adaptation across diverse product categories may require tuning.
Hybrid Models	Balanced performance by combining decomposition and ML models [4].	Depends on integration of multiple data sources.	Real-time capability depends on pipeline optimization.	Increased complexity may affect large-scale deployment.

the perspective of pharmaceutical sales prediction. This review establishes that, indeed, machine learning algorithms such as XGBoost and GRNN are more accurate, scalable, and flexible than traditional statistical methods. Additionally, it is possible to create a more holistic approach by incorporating auto-preprocessing [12], temporal feature engineering, and hybrid model integration.

Despite these promising findings, there are a few limitations and scope for improvement. For instance, though promising, there is scope for improvement and validation for highly volatile datasets, improving the interpretability of complex models, and addressing issues related to incorporating external covariates such as supply chain disruptions, demographic changes, and policy-driven demand changes [1], [3]. Future research directions and scope for improvement would involve increasing datasets, improving model explainability using interpretable boosting techniques, and validating the framework in live situations.

To conclude, it is evident that the application of machine learning-based forecasting presents a revolutionary prospect for the pharmaceutical industry as a whole. The integration of robust preprocessing techniques, efficient modeling approaches, and a deployment-ready dashboard represents a highly promising approach in terms of ensuring enhanced reliability in forecasting outcomes and facilitating transparency and informed decision-making for stakeholders involved. The application of machine learning-based forecasting is likely to significantly improve the resilience of the pharmaceutical supply chain.

REFERENCES

[1] H. Ahaggach, L. Abrouk, and E. Lebon, "Systematic Mapping Study of Sales Forecasting: Methods, Trends, and Future Directions," *Forecasting*, vol. 6, no. 3, pp. 502–532, 2024. doi:10.3390/forecast6030028.

[2] L. Hobor, M. Brcic, L. Polutnik, and A. Kapetanovic, "Comparative Analysis of Modern Machine Learning Models for Retail Sales Forecasting," *arXiv preprint* arXiv:2506.05941, 2025. doi:10.48550/arXiv.2506.05941.

[3] O. O. Mustapha and D. T. Sithole, "Forecasting Retail Sales using Machine Learning Models," *American Journal of Statistics and Actuarial Sciences*, vol. 6, no. 1, pp. 35–67, 2025. doi:10.47672/ajsas.2679.

[4] C. S. Bojer, "Understanding machine learning-based forecasting methods: A decomposition framework and research opportunities," *International Journal of Forecasting*, vol. 38, no. 4, pp. 1555–1561, 2022. doi:10.1016/j.ijforecast.2021.11.003.

[5] Y. I. Akande, J. Misra, A. Misra, S. Misra, O. O. Akande, and R. Ahuja, "Application of XGBoost Algorithm for Sales Forecasting Using Walmart Dataset," in *Lecture Notes in Networks and Systems*, Springer, 2022. doi:10.1007/978-981-19-1111-8_13.

[6] C. Neba, S. F. Chenwi, G. Nsuh, G. Agbara, P. Neba, A. Webnda, F. Ikpe, V. Orelaja, A. Sylla, and Nabintou, "Advancing Retail Predictions: Integrating Diverse Machine Learning Models for Accurate Walmart Sales Forecasting," *Asian Journal of Probability and Statistics*, vol. 26, pp. 1–23, 2024. doi:10.9734/ajpas/2024/v26i7626.

[7] D. Xie and Z. Shilong, "Machine Learning Model for Sales Forecasting by Using XGBoost," in *Proc. ICCECE*, pp. 480–483, 2021. doi:10.1109/ICCECE51280.2021.9342304.

[8] S. H. U. H. Andrabi, K. Alice, and S. A. Srivastava, "Sales Forecasting using XGBoost," *TechRxiv Preprint*, 2022. doi:10.36227/techrxiv.21444129.v1.

[9] A. Tawakuli, B. Havers-Zulka, V. Gulisano, and D. Kaiser, "Time-Series Data Preprocessing: A Survey and an Empirical Analysis," *Journal of Engineering Research*, 2024. doi:10.1016/j.jer.2024.02.018.

[10] M. Usmani, Z. Memon, A. Zulfiqar, and R. Qureshi, "Preptimize: Automation of Time Series Data Preprocessing and Forecasting," *Algorithms*, vol. 17, no. 8, p. 332, 2024. doi:10.3390/a17080332.

[11] K. Fourkiotis and A. Tsadiras, "Applying Machine Learning and Statistical Forecasting Methods for Enhancing Pharmaceutical Sales Predictions," *Forecasting*, vol. 6, pp. 170–186, 2024. doi:10.3390/forecast6010010.

[12] Q. Hidayaturrohman and E. Hanada, "Impact of Data Pre-Processing Techniques on XGBoost Model Performance for Predicting All-Cause Readmission and Mortality Among Patients with Heart Failure," *BioMedInformatics*, vol. 4, pp. 2201–2212, 2024. doi:10.3390/biomedinformatics4040118.

[13] C. Deb, F. Zhang, J. Yang, S. Lee, and K. Shah, "A review on time series forecasting techniques for building energy consumption," *Renewable and Sustainable Energy Reviews*, vol. 74, pp. 902–924, 2017.

[14] K. Kim, "Financial time series forecasting using support vector machines," *Neurocomputing*, vol. 55, no. 1–2, pp. 307–319, 2003.

[15] G. Candan, M. Taskin, and H. R. Yazgan, "Demand forecasting in pharmaceutical industry using neuro-fuzzy approach," *J. Mil. Inf. Sci.*, vol. 2, pp. 41, 2014.
[16] A. Jemal, E. Ward, Y. Hao, and M. Thun, "Trends in the leading causes of death in the United States, 1970–2002," *JAMA*, vol. 294, no. 10, pp. 1255–1259, 2005.
[17] G. Dudek, "Neural networks for pattern-based short-term load forecasting: A comparative study," *Neurocomputing*, vol. 205, pp. 64–74, 2016.
[18] J. Ord, R. Fildes, and N. Kourentzes, *Principles of Business Forecasting*, 2nd ed., Wessex Inc., 2017.
[19] S. Makridakis, A. Andersen, R. Carbone, R. Fildes, M. Hibon, R. Lewandowski, J. Newton, E. Parzen, and R. Winkler, "Accuracy of forecasting: An empirical investigation," *Journal of the Royal Statistical Society. Series A (General)*, vol. 142, no. 2, pp. 97–145, 1983. doi:10.2307/2982005.
[20] S. B. Jabeur, S. Mefteh-Wali, and J. L. Viviani, "Forecasting gold price with the XGBoost algorithm and SHAP interaction values," *Annals of Operations Research*, vol. 334, pp. 679–699, 2021. doi:10.1007/s10479-019-03377-6.