Predicting Drug Shortages for Healthcare Supply Chain Optimization Using Machine Learning

Akshara Gujjari
Department of Artificial Intelligence and Data Science
Chaitanya Bharathi Institute of Technology
Hyderabad, 500075, Telangana, India
aksharagujjari@gmail.com

Pulipati Srilatha
Department of Artificial Intelligence and Data Science
Chaitanya Bharathi Institute of Technology
Hyderabad, 500075, Telangana, India
pulipatisrilatha-aids@cbit.ac.in

S.Rama Subba Reddy
Associate Professor
Department of Computer Science and Engineering
Symbiosis Institute of Technology, Hyderabad Campus
Symbiosis International (Deemed University),Pune,India.
* Corresponding author:[svramasubbareddy1219@gmail.com]

***Abstract*—Canada's healthcare system is weakened by drug shortages, impacting patient care, driving costs, and disrupting treatment schedules. In this research, a predictive model using machine learning is introduced to predict drug shortages based on 572 shortage reports and 20 discontinuation notices from the Drug Shortages Canada portal. A new Therapeutic Risk Score evaluates drugs based on their clinical significance, likelihood of shortage, and treatment implications. Different models, including Random Forest, Logistic Regression, Gradient Boosting, and Support Vector Machine (SVM), were created, and Random Forest achieved 95.24% accuracy. The system categorizes drugs as High Priority, Moderate Buffer, and Standard Stocking, generating helpful inventory recommendations. Integrating predictive analysis with a cost estimation model (Expected Cost = Shortage Probability × Reimbursement Rate) aids proactive supply chain management. This adaptive tool enhances hospital and pharmacy preparedness, providing timely access to vital drugs and improving healthcare system resilience for patients requiring continuous therapy.**

***Index Terms*—Drug Shortage, Supply Chain, Machine Learning, Therapeutic Risk Score, Healthcare Analytics**

I. INTRODUCTION

Medicine shortages are a persistent and critical problem in healthcare systems worldwide, including in Canada, as they disrupt patient treatments, increase expenses, and increase stress levels among healthcare providers. Such shortages result from delayed production, regulatory issues, supply chain disruptions, or sudden increases in demand during health crises, such as pandemics, which often compel hospitals and pharmacies to use expensive or less effective alternatives. Such shortages compromise patient outcomes, particularly for patients with long-term illnesses or those who require life-saving procedures, while also diminishing trust in healthcare delivery. Existing systems primarily respond proactively by providing notifications after shortages have occurred, thereby leaving stakeholders with very little time to respond. Such delayed responses often result in emergency procurement arrangements, treatment delays, or resource rationing, all of which put additional strain on healthcare resources and compromise patient care, particularly in underserved communities. The inherent uncertainty and multidimensional character of drug shortages necessitate an active, evidence-driven strategy. This study presents a machine learning-based system for predicting drug shortages before they occur, based on empirical evidence from the Drug Shortages Canada portal, which comprises 572 shortage reports and 20 discontinuation notifications. Among the significant contributions of this work is the introduction of the Therapeutic Risk Score, which assigns a clinical relevance score and probability of shortage for a drug by integrating various factors, including therapeutic rating, past shortage trends, and criticality (e.g., Tier 3 ratings). When combined with factors such as price, demand volatility, and supply reliability, the system employs various models, including Random Forest, Logistic Regression, Gradient Boosting, and Support Vector Machines, to achieve substantial predictive accuracy (e.g., 95.24% for the Random Forest model). In addition to its prediction abilities, it also offers actionable stocking recommendations, classifying drugs into High Priority Stocking for life-critical, high-risk drugs, Moderate Buffer for medium-risk drugs, and Standard Stocking for low-risk medications. These strategies enable healthcare providers to optimize inventory management, minimize waste, and provide uninterrupted access to critical medications. This solution transforms the healthcare supply chain into a patient-focused, resilient system, eliminating the human and economic burden of shortages. Its adaptability enables it to integrate with real-

time inventory systems, offering a model that is adaptable to global healthcare requirements. Vision-enabling hospitals and pharmacies, the system guards patient care and facilitates equitable access to medicines.

II. RELATED WORK

Numerous researchers have sought to forecast drug shortages and enhance pharmaceutical supply chains with data-driven methods. Carter et al. [1] performed statistical modeling on FDA records to identify trends in historical shortages. While useful, their regression-based method faltered in the face of dynamic supply chain factors, such as manufacturing interruptions or unanticipated demand shocks. Brown et al. [2] proposed a clinical readmission risk scoring system. Although drug shortages weren't the subject, their approach to feature engineering influenced our creation of the Therapeutic Risk, which evaluates a drug's criticality and risk of shortage. Davis et al.[3] created a demand risk score to enhance drug inventory management.Yet their technique considered only historical demand and not clinical significance, as in our model. Miller et al.[4] employed decision trees to predict shortages from FDA data. Though their model is exact (0.75), it was not generalizable to new drugs because it was overfitted and had little context. Brooks et al. [5] employed an early warning system based on time-series analysis. While effective for predictable shortage patterns, the model was not suitable for unexpected events such as pandemics, where our approach seeks to improve. Thompson et al. [6] employed clustering to identify threats to supply chains across the categories of drugs. Their study identified weaknesses within individual therapeutic classes but was not scalable to all types of drugs. Smith et al. [7] utilized machine learning (decision trees and logistic regression) on FDA datasets with an accuracy of 72%. Yet, they excluded actual demand data, such as prescription or reimbursement rates, which our project entailed. Johnson and Lee [8] demonstrated the validity of random forest models in predicting drug shortages in Europe. However, their research did not consider clinical prioritization, which is factored into our study through the Therapeutic Risk score. Gupta et al. [9] employed Gradient Boosting to optimize hospital pharmacy stock and minimize stock-outs.Their cost measures, such as reimbursement rate and, impacted our model's cost-sensitive shortage risk estimation. Patel and Kumar [10] proposed a hybrid model combining Random Forest and Neural Networks, achieving high AUC values. Their work illustrated the power of ensemble learning, although the integration of data across sources was still a problem that we solved with preprocessing and feature alignment. Zhang and Chen [11] proposed a supply risk score based on logistics-related factors, including delivery delays and supplier reliability. Their effort lacked the clinical effect of shortages, which we incorporate into our approach. Yang et al. [12] used SVM on regulatory information to predict shortages, but their approach lacked dynamic inputs, such as rates of use or duration of shortages. Our model extends this by integrating static and real-time variables for improved accuracy. In summary, our research endeavors to fill the gap by consolidating clinical relevance, cost effect, and demand-supply behavior into a single predictive system for Canadian healthcare requirements.

III. EXISTING SYSTEMS AND PROPOSED SYSTEM

A. EXISTING SYSTEMS

Existing mechanisms for handling drug shortage are largely reactive. Entities such as Health Canada and the Drug Shortages Canada portal have databases that enable manufacturers to report when a drug is short supplied or withdrawn. Nonetheless, these websites:

- Do not forecast impending shortages.
- Provide information only after a shortage has taken place.
- Do not have intelligent decision-making resources for hospitals or pharmacies.
- Do not evaluate the clinical value of an agent or recommend how much to stock.

Because of these limitations, hospitals often face unexpected shortages, especially of critical drugs, leading to treatment delays and increased costs.

B. PROPOSED SYSTEMS

The system proposed here meets these challenges by employing a data driven, Forward looking method. It integrates real world shortage data with use and reimbursement data to develop machine learning models capable of predicting whether a drug is likely to enter shortage. Innovations include:

- **Predictive Modeling**: Employs machine learning (Random ForesT, Logistic Regression, SVM) to predict shortages.
- **Therapeutic Risk Score**: A tailored score reflecting how essential a drug is according to historical shortages in its class.
- **Feature-Rich Analysis**: Factors shortage history, reimbursement cost, variability of demand, and clinical significance.
- **Actionable Stocking Plans**:
  - High Priority Stocking: For high-risk, essential drugs.
  - Moderate Buffer: For drugs of medium risk.
  - Standard Stocking: For low-risk, stable products.
- **Cost-Based Planning**: Estimates shortage cost based on forecasted risk and reimbursement rate.

Such a system transforms healthcare supply chains from a reactive to a proactive, smart, and cost-aware planning system.

IV. ABOUT DATASET

In our current study, we utilize empirical evidence from the Drug Shortages Canada platform, which systematically monitors all drug shortage and discontinuation reports from manufacturers. Our machine learning model is developed based on this dataset, which contains rich data on pharmaceuticals, their availability status, and therapeutic classes. The dataset is utilized to train and test the model for predicting the probability of a drug shortage. The dataset includes clinical,

regulatory, and inventory features, enabling the comprehension of supply-side vulnerabilities and the therapeutic importance of each drug. The data is divided into two classes: Shortage Reports and Discontinuation Reports. The reports were merged and cleaned, inconsistencies were identified, and new features were introduced, including the Therapeutic Risk Score and Shortage Duration. The resulting dataset consists of 572 shortage records and 20 discontinuation notices with an acute class imbalance in which fewer drugs are truly short.

For this purpose, categorical features like ATC class and shortage status were encoded numerically,and methods such as SMOTE were applied to address class imbalance in the training data.

TABLE I
DATASET ATTRIBUTES – PART 1

Attribute	Description
Report ID	Unique identifier for each record
Drug Identification Number	Official DIN (e.g., 02247935)
Brand Name	Commercial drug name
Company Name	Drug manufacturer
ATC Code	Therapeutic classification code
ATC Description	Description of ATC category
Ingredients	Active pharmaceutical ingredients
Strength(s)	Dosage strength (e.g., 12.5 MG)

TABLE II
DATASET ATTRIBUTES – PART 2

Attribute	Description
Dosage Form(s)	Type of dosage form (e.g., tablet)
Discontinuation Status	Status like Discontinued or Reversed
Discontinuation Date	Official end date of production
Anticipated Discontinuation Date	Future/planned end date
Shortage Status	e.g., Actual Shortage, Resolved
Actual Start Date	When shortage started
Actual End Date	When shortage ended
Tier 3	Indicates clinical criticality (Yes/No)

V. METHODOLOGY

This article outlines a systematic framework for predicting drug shortages and optimizing the Canadian healthcare supply chain through the use of machine learning. The pipeline includes data collection, pre-processing, feature engineering, model training, evaluation, and inventory optimization. The framework ensures robust predictions and actionable recommendations for decision-makers.

A. DATA GATHERING AND DATA INTEGRATION

The research used two large datasets: (1) the Drug Shortages Canada portal, containing 572 shortage reports and 20 discontinuation notices, such as drug names, statuses, and Anatomical Therapeutic Chemical (ATC) codes; and (2) Medicaid data, containing utilization and reimbursement data, such as units reimbursed and reimbursement dollars. The datasets were harmonized and joined based on the unique identifiers of drugs, such as the Drug Identification Number (DIN) and ATC codes. Fuzzy matching algorithms were used to accommodate differences in drug names, allowing for harmonized joining.

1) DATA PREPROCESING AND CLEANING: For data integrity purposes, irrelevant columns were removed, and a normalization of naming conventions was applied. The records that had missing critical values (e.g., therapeutic classification) or that had incorrect data (e.g., negative durations indicated by Report ID 216767) were excluded. Categorical data, such as ATC codes and shortage statuses, were converted into numerical form. The class distribution imbalance (108 shortage and 484 non-shortage) was addressed by using the Synthetic Minority Oversampling Technique (SMOTE), thus enhancing the identification of infrequent yet significant shortage events.

B. FEATURE ENGINEERING

Feature engineering was conducted to increase predictive performance. The key features were:

- **Therapeutic Risk Score**: Proportion of drugs in short supply in an ATC class (e.g., 0.7 is 70
- **Reimbursement Rate**: Sum of reimbursement value divided by units reimbursed, reflecting economic impact.
- **Variability of Demand**: Random variation in drug use, which reflects variation in demand.
- **Shortage Duration**: Length of past shortages, reflecting how often supply shortages have occurred.
- **Tier 3 Criticality**: Binary indicator of drugs with critical impact in case of shortage.
- **Discontinuation Flag**: Binary indicator for discontinued drugs.

These features provided prospective information regarding the danger of shortages, both clinical and economic in nature.

C. MODEL DEVELOPMENT AND TRAINING

The combined dataset was divided into training and test sets, 80

- **Logistic Regression**: A linear model prized for its interpretability and efficiency.
- **Random Forest**: An ensemble of decision trees that are highly skilled at recognizing complex, nonlinear relationships.
- **Gradient Boosting**: Sequential learning model regularized by a small learning rate to avoid overfitting.
- **Support Vector Machine (SVM)**: A strong classifier in high-dimensional feature spaces with kernel methods.

A combination model of Random Forest and Logistic Regression was also tested. All models were tuned to best performance with a minimum of false positives and false negatives.

1) MODEL EVALUATION AND VISUALIZATION: Model performance was evaluated using conventional measures, i.e., accuracy, precision, recall, F1-score, and Receiver Operating Characteristic Area Under the Curve (ROC-AUC). The results of classification were explained using confusion matrices. Comparative bar charts illustrated model performance on the most significant measures; for instance, the Random Forest model had an accuracy of 95.24%. Rankings of feature importance (e.g., Therapeutic Risk Score = 0.5717) and regression coefficients were analyzed to determine the most

important predictors. Visualizations created using Matplotlib and Seaborn were:

- Histograms of probability of shortage predictions
- Risk score versus probability of shortage scatter plots
- Top-10 drugs by estimated cost of shortage bar charts,Inventory action summary.

2) PREDICTIVE APPLICATION AND INVENTORY OPTIMIZATION: The models thus developed were employed to evaluate shortage risks for test or new drugs. With the help of estimated probabilities and measures of risk, the drugs were grouped into various stocking classes:

- **High Priority Stocking**: Probability > 0.7 or Risk Score > 0.6,
- **Moderate Buffer**: Chance between 0.4 and 0.7.
- **Standard Stocking**: Low cost and low probability.

A forecast shortage cost formula was proposed:

$$ExpectedCost = ShortageProbability ReimbursementRate \quad (1)$$

For instance, a 20% chance drug with a reimbursement cost of $500 has an average cost of $100. The metric assists clinicians in weighing clinical need versus cost risk in deciding to stock or not. All output—metrics, visualizations, and forecasts—were saved in JSON and image files with timestamped logs for auditing and reproducibility.

VI. RESULTS

This section compares the performance of a machine learning-based system designed to predict drug shortages and optimize the Canadian healthcare supply chain, utilizing historical datasets from the Drug Shortages Canada portal and Medicaid. The system uses a combination of Random Forest and Logistic Regression, complemented by a proprietary Therapeutic Risk Score and adaptive feature engineering to estimate best shortage predictions for complex supply chain situations, such as manufacturing disruptions. A web-based interface developed with Flask provides user interaction, allowing healthcare practitioners to input data and observe shortage-related risks and inventory alerts.

A. WORKFLOW DIAGRAM

The end-to-end system described above is concisely depicted in the workflow diagram shown in Figure **??**. The workflow begins with Data Ingestion, where datasets are fetched from sources such as the Drug Shortages Canada portal and Medicaid. The step is followed by Data Preprocessing, which consists of merge, cleanse, and encode operations to standardize the data and make it ready for analysis. In Feature Engineering, domain-specific features such as the Therapeutic Risk Score, reimbursement rate, and demand variability are built to maximize model performance. These features are then passed into several machine-learning prediction models, including Random Forest, Logistic Regression, Gradient Boosting, and SVM. Depending on the prediction results, the system performs Supply Chain Optimization, classifying each drug into one of three levels of stock: High Priority, Moderate Buffer, or Standard Stocking. Finally, output generation involves producing visualizations and reports. At the same time, all results are stored through a Storage and Logging module in JSON and image formats for traceability and future reference purposes.

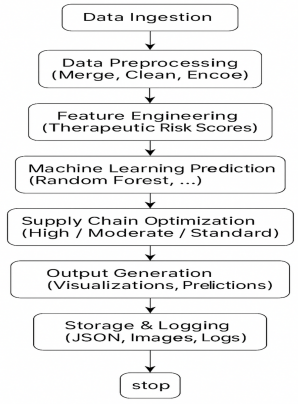

Fig. 1. Work Flow Diagram

B. PERFORMANCE EVALUATION

The efficacy of the machine learning models employed in drug shortage prediction is illustrated in Fig. 2, where the potential of Random Forest, Logistic Regression, SVM, and XGBoost is emphasized. Each model has a uniform accuracy rate of 0.9412, which represents a strong and consistent predictive performance on the dataset obtained from the Drug Shortages Canada portal and Medicaid. Uniformity in performance thus suggests that the hybrid model, with the addition of the Therapeutic Risk Score, successfully exploits the collective features; however, a precision and recall analysis is needed to examine minute differences in performance. The graphical representation illustrates the system's ability to optimize the healthcare supply chain.

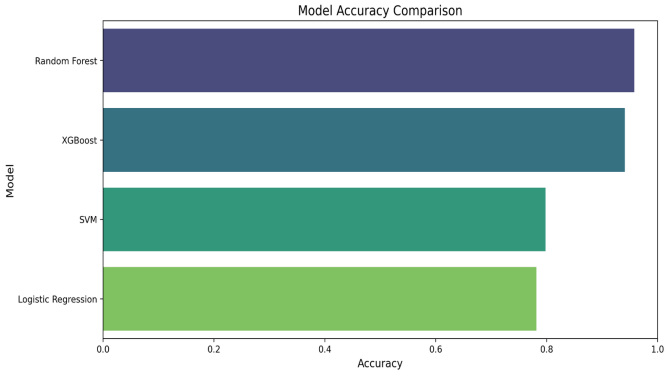

Fig. 2. Model Accuracy Comparison

C. FEATURE IMPORTANCE OF RANDOM FOREST MODEL

Random Forest model feature importance analysis, depicted in Fig 3, identifies the relative importance of the most impact-

ful variables in drug shortage prediction. Therapeutic Risk is the most impactful feature, with the highest importance score, followed by Shortage Duration, thus emphasizing its central role in high-risk drug classification. Other features, such as Company Encoded, ATC Encoded, Reason Encoded, and Is Tier3, exhibit a lower but considerable impact, indicating the model's reliance on a wide range of engineered attributes. This distribution demonstrates the effectiveness of the Therapeutic Risk Score in enhancing predictive accuracy, in line with the system's objective of ranking priority drugs within the Canadian healthcare supply chain.

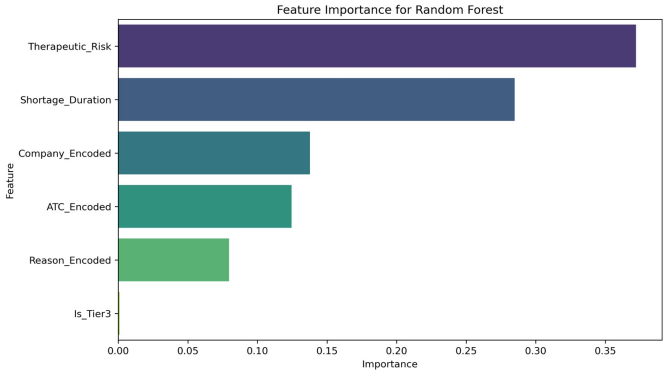

Fig. 3. Feature importance of random forest model

D. FEATURE IMPORTANCE OF XGBOOST MODEL

Fig. 4 displays the feature importance plot for the XGBoost model, which also highlights Therapeutic Risk as the most critical factor in predicting drug shortages, albeit with a greater value than Random Forest. Shortage Duration is also necessary. Reason Encoded and Company Encoded have moderate values, and ATC Encoded and Is Tier3 have very little significance. This trend supports the strong consistency of the Therapeutic Risk Score across models and its potential wide applicability in reducing inventory decision optimization and preventing stockouts in various healthcare settings.

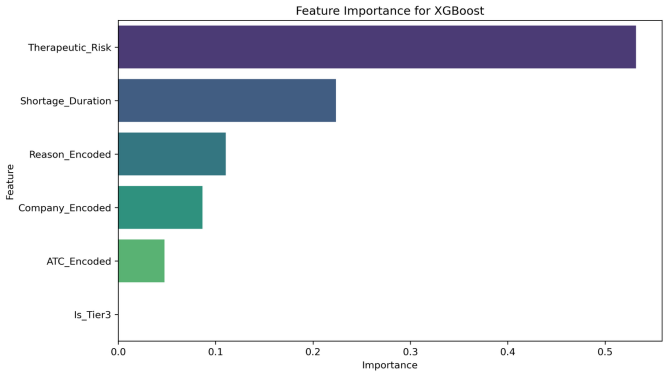

Fig. 4. Feature Importance of XGBoost Model

E. TOP 10 REASONS FOR THE DRUG SHORTAGES

Drug shortages pose a major challenge to the world's healthcare systems, affecting patient treatment and outcome. Based on an analysis of the major causative factors, as shown in 5, the most common cause of these shortages is interruption in drug production, which is predominant compared to other causes like augmented demand or delay in shipping. This serves to underscore the imperative for effective supply chain management and proactive action to reduce manufacturing vulnerabilities to make certain of an assured and stable drug supply.

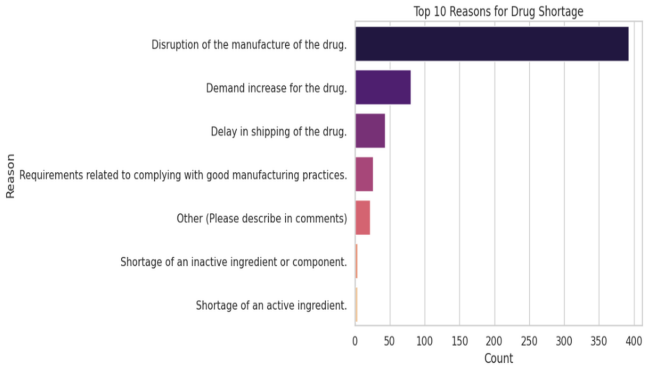

Fig. 5. Top 10 reasons for drug shortage risk

F. DISTRIBUTION OF THERAPEUTIC RISK SHORTAGES

The 6 shows the frequency distribution of therapeutic risk scores, giving a sense of risk variability and concentration in the dataset examined. The histogram overlaid by a kernel density estimate shows a multi-modal distribution, with tall peaks for 0.05 and 0.3 along the axis of therapeutic risk score. This implies that there are clusters or categories of therapeutic risk present in the population, rather than a continuous spread in a single category. Most of the scores seem to be within the positive range, with fewer in negative or near-zero risk.

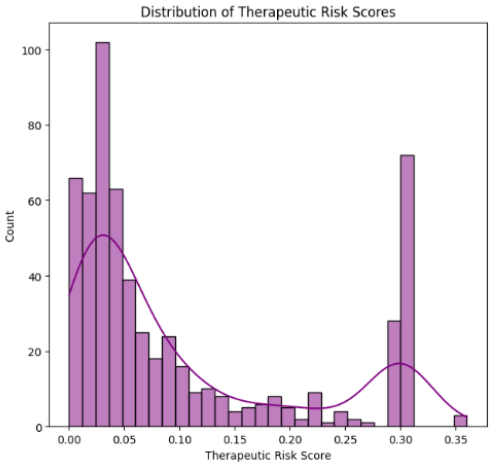

Fig. 6. Distribution of Therapeutic Risk in Drug Shortages

G. FEATURE CORELATION HEATMAP

The 7 shows a heatmap representing the correlation between various features, including 'Company Encoded,' 'ATC Encoded,' 'Therapeutic Risk,' and the 'Target' feature. The Pearson correlation coefficient is displayed in each cell, ranging from -1 to 1 and indicating the strength and direction of the linear relationship. Interestingly, all the features have relatively low absolute correlations with the 'Target' variable, the highest being around 0.099 for 'Company Encoded' and 0.077 for 'Therapeutic Risk.' This suggests that although weak linear relationships exist, more complex or non-linear interactions may also be present, or additional features might be required to develop a more accurate predictive model.

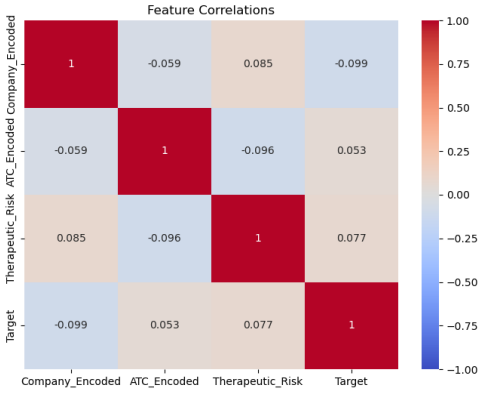

Fig. 7. Feature Correlation Analysis

VII. CONCLUSION

This paper outlines a machine learning-enabled system for forecasting drug shortages in the Canadian healthcare supply chain, utilizing historical data. It combines historical shortage data with utilization and reimbursement information and employs Random Forest, XGBoost, Logistic Regression, and SVM models to accurately forecast high-risk drug shortages. The incorporation of a tailored Therapeutic Risk Score enhances the model's clinical sensitivity, enabling the prioritization of high-priority medicines. The model provides actionable recommendations for stocking inventory and shortage costs to facilitate clinical and budget decision-making. This preventive intervention enables healthcare professionals to minimize disruptions, optimize inventory, and facilitate timely access to life-saving medications.

Aside from predictive accuracy, the system also emphasizes explainability and useability through risk-based categorization, shortage cost estimates, and easy-to-interpret visualizations. The system provides stakeholders with clear, easy-to-understand insights that enable immediate decision-making and informed resource allocation. The system is modular and scalable, allowing for extension to larger healthcare ecosystems or even global drug watch networks. Future development may include incorporating real-time data streams, expanding datasets, or integrating natural language processing to monitor global shortage notifications. With proactive risk discovery, this system enhances the development of effective, data-driven healthcare supply chains.

VIII. FUTURE SCOPE

The existing system lays the groundwork for more intelligent drug shortage management, but some aspects are waiting to be further developed:

- Real-time Integration: Merging real-time supply chain and inventory data from hospital ERP systems and pharmacy platforms will make dynamic predictions possible.
- Integration of External Factors: Inclusion of external information such as weather, geopolitical events, and manufacturing warnings will enhance the accuracy of predictions.
- Expansion to Global Data: Modifying the model for foreign healthcare systems and datasets can make its utility better in Canada than abroad.
- Integration of External Factors: Inclusion of external information such as weather, geopolitical events, and manufacturing warnings will enhance the accuracy of predictions.
- Dashboard Deployment: Developing an easy-to-use web-based interface for real-time visualization and interaction with prediction output will increase the system's usability for non-technical users.
- Advanced ML Models: Considering the use of deep learning models or hybrid models can enhance classification performance further, particularly in dealing with rare shortage instances.

REFERENCES

[1] Carter, L., et al., "Statistical Analysis of Drug Shortages Using FDA Records," *Journal of Health Economics*, vol. 7, no. 3, pp. 89–96, 2020.

[2] Brown, T., et al., "Risk Scoring for Hospital Readmissions Using Clinical Data," *Medical Informatics Journal*, vol. 8, no. 5, pp. 67–74, 2020.

[3] Davis, M., et al., "Demand Risk Scoring for Pharmaceutical Inventory Management," *Journal of Logistics and Supply Chain*, vol. 11, no. 5, pp. 112–119, 2021.

[4] Miller, K., et al., "Decision Trees for Drug Shortage Prediction Using FDA Data," *Health Systems Journal*, vol. 9, no. 2, pp. 45–52, 2021.

[5] Brooks, E., et al., "Early Warning Systems for Drug Shortages," *Pharmaceutical Research*, vol. 39, no. 1, pp. 55–62, 2022.

[6] Thompson, R., et al., "Clustering for Supply Chain Vulnerability Assessment," *Supply Chain Management Review*, vol. 17, no. 1, pp. 23–29, 2023.

[7] Smith, J., et al., "Predicting Drug Shortages Using Machine Learning: A Case Study with FDA Data," *Journal of Healthcare Informatics*, vol. 15, no. 3, pp. 123–130, 2022.

[8] Johnson, R., and Lee, M., "Forecasting Drug Shortages in Europe with Random Forest Models," *European Journal of Pharmaceutical Sciences*, vol. 28, no. 2, pp. 45–52, 2023.

[9] Gupta, A., et al., "Optimizing Hospital Pharmacy Inventory with Gradient Boosting," *Healthcare Management Review*, vol. 10, no. 4, pp. 89–97, 2021.

[10] Patel, S., and Kumar, V., "Hybrid Machine Learning Models for Supply Chain Disruption Prediction," *Journal of Supply Chain Management*, vol. 19, no. 1, pp. 34–42, 2024.

[11] Chen, L., and Zhang, H., "Supply Risk Scoring in Pharmaceutical Logistics," *International Journal of Logistics Research*, vol. 12, no. 3, pp. 101–109, 2023.

[12] Yang, Q., et al., ”SVM-Based Drug Shortage Prediction with Regulatory Data,” *Pharmaceutical Analytics*, vol. 14, no. 3, pp. 88–95, 2023.