Development of an AI-Driven Model for Drug Sales Prediction Using Enhanced Golden Eagle Optimization and XGBoost Algorithm

Ying Xiong
Business School of Dongguan City University, Dongguan, Guangdong, 523000, China
E-mail: xiongying202304@163.com
*Corresponding author

Keywords: drug sales, enhanced golden eagle optimized extreme gradient boosting (EGEO-Xgboost) resource management

Received: November 4, 2024

The pharmaceutical industry plays a crucial role in public health by providing essential medications to address various medical conditions. Predicting drug sales accurately is paramount for pharmaceutical companies to efficiently manage their resources, plan production, and optimize marketing strategies. To address this an AI-based model for predicting medication product sales based on the Enhanced Golden Eagle Optimized and Extreme Gradient Boosting (EGEO-XGBoost) framework. The technique begins with data collecting from the Kaggle website, followed by pre-processing with Min-max normalization to remove noise and assure consistency. The preprocessed data is then used to extract relevant features via Linear Discriminant Analysis (LDA). The enhanced EGEO method fine-tunes the parameters of the XGBoost model, improving its predictive ability. The comparative findings demonstrate that the proposed method significantly improves accuracy (0.90), specificity (0.86), sensitivity (0.92), MCC (0.82), F1-score (0.94), and RMSE (4.02). Incorporating such predictive models into the decision-making processes of pharmaceutical corporations can result in improved resource management, better marketing plans, and increased operational effectiveness.

Povzetek: Predlagan je model za napovedovanje prodaje zdravil na osnovi izboljšanega algoritma Extreme Gradient Boosting, optimiziranega z metodo Enhanced Golden Eagle Optimization (EGEO-XGBoost).

1 Introduction

The pharmaceutical industry was critical to enhancing the world's health through the development and provision of life-saving drugs. Strategic planning by pharmaceutical companies is highly reliant on their ability to forecast the sales of drugs. They can use it to effectively manage resources, optimize production, and make decisions on marketing and sales plans with knowledge [1]. Traditional sales forecasting methods often yield suboptimal outcomes because they fail to capture the dynamic and complex nature of the pharmaceutical industry.

Some of the challenges with traditional methods include the incapability to efficiently process large and diversified datasets and, the failure to take into account extraneous factors like changes in regulatory laws, public health incidents, and macroeconomic conditions. All these shortcomings cause issues such as overstocking or understocking pharmaceuticals, missing business opportunities, and failing to respond on time to market changes in demand [2, 3].

The pharmaceutical sector has undergone huge changes over time due to multiple reasons, among which are increases in medical expertise, changing regulatory environments, shifting consumer habits, and varying healthcare policies. These dynamic ingredients create uncertainties and challenges that even traditional forecasting models are unable to address properly. Hence, pharmaceutical companies are often faced with issues such as overstock or understocking of pharmaceuticals, missing business opportunities, and not responding to the changing trend of demand [4].

To bridge these gaps, the interest in AI and ML in predictive models has generated much momentum. AI-based models allow efficient computations of large datasets of historical sales records, demographic information, prescription trends, and competitor analysis, altogether integrated with external factors for comprehensive forecasting. Such models differ from traditional techniques in their ability to pick up on some minute patterns and adapt according to market dynamics in real-time [5].

These AI-driven drug sales forecasting models come with multiple benefits. These include optimized production planning, reduced waste, and improved management of inventory. In addition, by aligning marketing strategies with promising markets and client segments, return on investment can be maximized while also supporting better resource allocation [6, 7]. Besides, the incorporation of external variables such as trends in public health and regulatory changes makes the prediction more accurate and holistic. However, challenges to implementing AI-based models abound, such as high-quality and updated data requirements, concerns regarding data privacy, regulatory compliance, and the interpretability of complex models [8]. Nonetheless, ongoing improvements in AI and ML techniques are poised to deliver transformational solutions in the sales forecasting space of the pharmaceutical sector.

By using AI-based predictive analytics, pharmaceutical companies will be able to achieve accurate, flexible, and data-driven forecasts. This will help ensure more efficient strategic planning and greater operational resilience in the long term. This may help to solve the dynamic nature of the industry while creating a more robust health ecosystem [9,10].

Aim of the research: An AI-based model to effectively predict drug sales by using Enhanced Golden Eagle Optimized and Extreme Gradient Boosting on the sales to ensure effective drug production. Improved resource management through data-driven techniques improves production planning and better strategies for marketing products.

1.1 Motivation of the research

The pharmaceutical industry is highly sensitive to the predictive inaccuracy of drug sales, an exercise that is important for maximal production, efficient distribution, and appropriate marketing strategy optimization. Traditional methods are inadequate in comparing the real complexity and dynamic nature of the market, and, consequently, resource allocations are suboptimal. Motivation is built on the interest in exploring how AI-based models can improve the forecasting of drug sales. The ability of large and diverse datasets, such as historical sales, market trends, and consumer demographics, to aid in the hidden patterns and adaptive change in conditions for AI systems helps pharmaceutical companies make more accurate, data-driven predictions and improve the management of their resources and accurately target specific customer segments, improving sales performance and access to lifesaving medications for patients.

1.2 Contribution

The following key points are described AI AI-based predicting drug sales contribution.

- **Data pre-processing with Min-Max normalization:** Aims to ensure balanced learning by normalizing heterogeneous pharmaceutical data into a similar range. It makes the model converge faster and prevents overfitting, resulting in more accurate and reliable predictions of drug sales.

- **Feature Selection via Linear Discriminant Analysis (LDA):** LDA particularly identifies and selects those features that have the most influence in preprocessed drug sales data. Focusing on these key features, LDA helps improve the accuracy of sales predictions, making sure that the model focuses on the most predictive factors for drug sales forecasting.

- **Combining XGBoost with Enhanced Golden Eagle Optimization (EGEO):** The model combines XGBoost, a powerful machine learning algorithm with the EGEO technique. This optimization approach fine-tunes the model's parameters to achieve higher accuracy in drug sales within the pharmaceutical business.

- **Performance evaluation:** It includes a thorough performance evaluation of the AI-based model by analyzing key metrics. These metrics establish the model's ability to provide accurate predictions on drug sales, which further enhances its practical use in the pharmaceutical industry for strategic decision-making purposes.

2 Related works

Table 1 shows demonstrates that the summary of previous studies.

Table 1: Summary of related works

Reference	Objective	Methods	Datasets	Key Findings
Bhattamisra et al., [11]	To explore AI applications in pharmaceutical and healthcare research.	Literature review, case studies on AI applications in healthcare.	PubMed, clinical trial data, AI healthcare applications.	AI plays a crucial role in enhancing drug discovery, personalized healthcare, and system optimization in the pharmaceutical industry.

Vora et al., [12]	To examine the role of AI in pharmaceutical technology and drug delivery design.	Review of AI-based drug delivery technologies and systems.	Pharmaceutical databases, drug delivery system data.	AI contributes to improving precision in drug formulation and the design of personalized drug delivery systems.
Shilong, [13]	To develop a machine learning model for sales forecasting using XGBoost.	XGBoost model for forecasting drug sales.	Drug sales data (historical sales data).	XGBoost effectively forecasts sales, showcasing high predictive accuracy for sales trends in the pharmaceutical sector.
Lin et al., [14]	To predict drug-drug interactions using multi-source data and transformer self-attention mechanism.	Multi-source data fusion, transformer self-attention mechanism.	Drug interaction databases.	The model successfully predicts drug-drug interactions with high accuracy, using advanced machine-learning techniques.
Tichy et al., [15]	To analyze trends in prescription drug expenditures.	Statistical analysis of prescription drug market data.	U.S. prescription drug expenditure data	Identifies trends in prescription drug costs, providing projections for future expenditures in healthcare systems.
Agrawal et al., [16]	To review AI applications in drug delivery systems and technology.	Case studies and literature review on AI-driven drug delivery.	Pharmaceutical and drug delivery system data.	AI enhances the precision of drug targeting, reduces side effects, and advances personalized medicine in drug delivery systems.
Biehn et al., [17]	To develop AI-based ADMET models for drug discovery.	ADMET modeling using AI-based platforms like SAFIRE.	SAFIRE platform, ADMET drug data.	Improves the accuracy of ADMET predictions, supporting faster and more efficient drug discovery processes.
Saikia et al., [18]	To review AI's role in therapeutic drug monitoring and clinical toxicity.	Literature review on AI in therapeutic monitoring.	Clinical therapeutic drug monitoring datasets, toxicology data.	AI plays an essential role in monitoring therapeutic drug levels, minimizing clinical toxicity, and optimizing drug safety.
Ali & Alrobaian, [19]	To evaluate the current and	Literature review and analysis of AI	Industry reports, and pharmaceutical	AI offers promising advancements in

	prospects of AI in pharmaceutical development.	technologies in pharmaceutical development.	development databases.	pharmaceutical product development, though challenges remain regarding large-scale implementation and real-world applications.
Łapińska et al., [20]	To introduce an AI-based application (SerotoninAI) for drug discovery.	AI-driven drug discovery platform focusing on serotoninergic systems.	Serotonergic drug data	AI-based SerotoninAI system enhances drug discovery, particularly for serotonergic systems, improving discovery times and accuracy.
Serrano et al., [21]	To assess AI applications in drug discovery and personalized medicine.	Review of AI applications in drug discovery and personalized drug delivery.	Pharmaceutical and clinical data, drug discovery databases.	AI is revolutionizing drug discovery, particularly in personalized medicine, by enabling more accurate and efficient drug delivery methods.
Jena et al., [22]	To explore AI-driven drug delivery systems in the pharmaceutical industry.	Literature review on AI's impact on drug delivery systems.	Industry and pharmaceutical data.	AI enhances the efficiency of drug delivery systems, particularly in designing customized delivery for patient-specific needs.
Visan & Negut, [23]	To explore AI integration in drug discovery and delivery systems.	Literature review on AI tools in drug discovery and delivery.	Pharmaceutical and drug discovery databases.	AI is transforming drug discovery and delivery, though its implementation still faces challenges in validation across diverse real-world applications.
Rathipriya et al., [24]	To validate various shallow and deep neural network methods for demand forecasting in the pharmaceutical industry.	Shallow and deep neural networks, RMSE for predictive accuracy evaluation.	Sales and demand data of eight pharmaceutical product groups.	Shallow neural network-based DFMs achieved a lower mean RMSE (6.27), proving more effective in estimating future demand than deep neural networks.
Jaganathan et al., [25]	To develop quantitative	Machine learning algorithms, feature	A dataset of 1,253 drug compounds	The SVM-based classifier achieved

	structure-activity relationship models for predicting drug-induced liver toxicity using machine learning.	selection techniques, support vector machines, 10-fold cross-validation.	with molecular descriptors.	an accuracy of 0.811, sensitivity of 0.840, specificity of 0.783, and MCC of 0.623, outperforming prior models in both internal and external validation.

2.1 Research gap

Existing methods applied in drug sales forecasting and drug discovery are very limited in the sense that their generalizability, scalability, and performance would be compromised to some extent in generalizing different drug classes, datasets, and regions. Our proposed model for predicting sales of pharmaceutical products, based on the Enhanced Golden Eagle Optimization and Extreme Gradient Boosting (EGEO-XGBoost) framework, fills these gaps by combining optimized techniques with a state-of-the-art predictive model for improved accuracy and scalability and improved generalization through adaptation across heterogeneous datasets and changing conditions, in turn providing even more reliable, actionable sales forecast in the pharma industry.

3 Methodology

A proposed methodology that employs an AI-based model for the forecasting of sales on pharmaceutical products uses a combination of an enhanced golden eagle optimized Extreme gradient boosting (EGEO-XGBoost). First, the data from Kaggle are preprocessed using Min-max normalization to eliminate noises and ensure consistency in data. A linear discriminate analysis technique is used to extract the features of the preprocessed data. This integrated approach will help enhance the accuracy of drug sales forecasting which can be used for better resource management and decision-making in the industry. Figure 1 illustrates the overview of the research framework.

3.1 Data gathering

The dataset gathered from kaggle [https://www.kaggle.com/code/milanzdravkovic/pharma-sales-data-analysis-and-forecasting/data], includes the drug sales collected across different stores over time. The main attributes included in the data set are the store ID, product ID, quantity of sales, and the date of transactions. Such time-series data is beneficial in the overall understanding of the performance of sales as well as analysis regarding trends, seasonality, and patterns. Data structured to accommodate sales forecasting, inventory planning, and business decisions related to pharmaceutical businesses. The collection method is primarily based on real-world sales activities so that the dataset obtained is relevant for predictive analytics and machine learning model development.

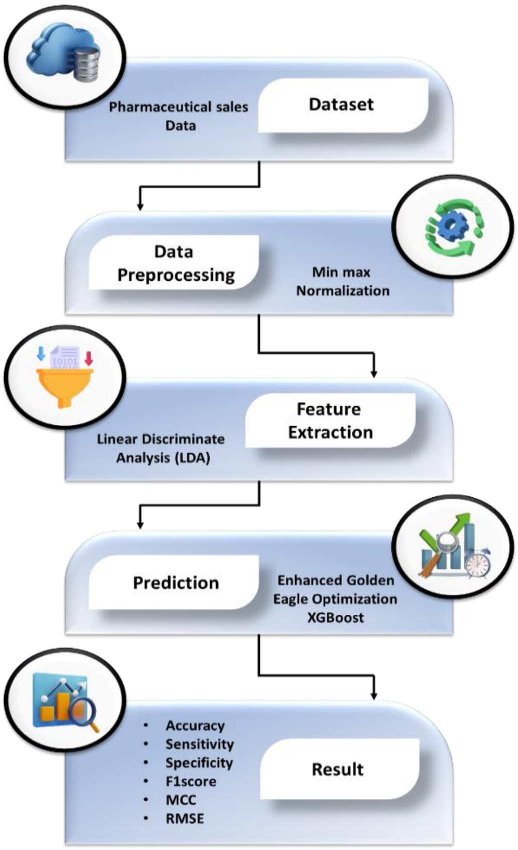

Figure 1: Overview of research to predict pharmaceutical product

3.2 Min-max normalization

A data preparation technique called min-max normalization, often referred to as feature scaling or min-max scaling, is used to change numerical data into a certain range, typically [0, 1]. With this normalization technique, all the numerical features in a dataset are put on an equal footing, making them comparable and preventing features with higher values from overpowering the analysis or adversely affecting machine learning algorithms. Using this technique, you can scale numerical data to a specific range. Data normalization or feature scaling are other names for it. Normalization aims to climb all of them evenly to make it easier to compare or understand a dataset's characteristics or variables. The following is the min-max normalization formula (1):

$$X_{new} = \frac{X - \min(X)}{\max(X) - \min(X)} \qquad (1)$$

Where $X_{new}$the result of normalizing is to produce a new value, X is the value of the former, $\max(X)$ is the highest value in the dataset, and $\min(X)$ is the smallest value in the dataset.

Before dividing by the range, which consists of the maximum and minimum values of the content, each data point must first be subtracted from the feature's minimum value. The importance of the dataset can be scaled between 0 and 1 following min-max normalization. Suppose the beginning values fall within a different ideal range (for instance, between -1 and 1). In that case, we may change the formula to reflect this by expressing the intended minimum and maximum values throughout the normalization process.

Even though min-max normalization was a straightforward and popular method, some datasets might have better solutions. Various scaling procedures, such as Z-score normalization or resilient scaling, may be more appropriate depending on the information's characteristics and the particular specifications of the machine learning method being applied.

3.3 Feature extraction

When forecasting Drug sales, the process of picking and developing pertinent features (variables) from the raw data that are likely to have a large impact on successfully predicting drug sales is referred to as feature extraction. By condensing the original data into a smaller representation while preserving the most crucial details, feature extraction aims to improve the accuracy of predictive models.

3.3.1 Linear discriminates analysis (LDA)

Linear discriminate analysis (LDA), a dimensionality-reduction and classification method, is used in statistics and machine learning. When the classes were distinct and evenly dispersed, it was extremely beneficial for addressing classification problems. The goal of LDA was to find a linear combination of characteristics that maximizes class separation while minimizing variation within each category. LDA's major objective was to maintain any information discriminating based on a person's class while projecting the original data into a lower-dimensional space. This was achieved by increasing the ratio of the within-class scatter matrix to the between-class scatter matrix.

The linear discriminate analysis (LDA) will be explained step-by-step using the following equations:

3.3.1.1 Making the mean vector calculations:

Assume there are m_u data points in each u class, which all exist. The mean vector for class u was calculated as the (equation 2) average of all feature vectors in class u, designated as m_{u}.

$$m_u = \frac{1}{n_u} \sum_{X \in \, classu} X \qquad (2)$$

3.3.1.2 Calculate the scatter matrix for the class (k_r):

The within-class scatter matrix k_r calculates the expansion or distribution of the data for each type. The scattered values from each class were added together to get the final result. The class's data points' covariance matrices are used as the scatter matrix for that class. The formula 3 for "k_r" is

$$k_r = \sum_{u=1}^{u} \sum_{X \in \, classu} (m_{u-m})(m_{u-m})^T \qquad (3)$$

3.3.1.3 Calculate the scatter matrix $\boldsymbol{k_b}$ between classes:

The spread or distribution of the class mean values was ascertained using the between-class scatter matrixk_b. The weighted sum of the outer products was used to determine the difference between the class means, and the overall mean across all data points."k_b" equation 4 is as follows:

$$k_b = \sum_{u=1}^{u} n_u (m_{u-m})(m_{u-m})^T \qquad (4)$$

Where

By dividing the average of all mc vectors by the sum of the data points in each class, the total mean vector of all the data points in the dataset was calculated using equation 5.

$$\mathrm{m} = \frac{1}{N} \sum_{u=1}^{u} n_u m_u \qquad (5)$$

3.3.1.4 Find the eigenvalues and eigenvectors for ($\boldsymbol{k_r^{-1}}$, $\boldsymbol{k_b}$):

The inverse of (k_b), written as ($k_r^{-1,}k_b$), is the product ($k_r^{-1,}k_b$), which we now compute. The product ($k_r^{-1,}k_b$)'s eigenvalues and associated eigenvectors v were then identified.

3.3.1.5 Find the highest (k) eigenvectors:

The eigenvalues are arranged in ascending order, and the top k eigenvectors are chosen to match the top k greatest eigenvalues. The transformation matrix R will comprise these "k" eigenvectors and project the data onto a lower-dimensional space.

3.3.1.6 Introduce the data into the new subspace:

The original data Y should be multiplied by the conversion matrix R produced in the previous phase to produce a new feature space with a lower dimensionality (equation 6).

$$[Y_{new}] = Y \cdot R \qquad (6)$$

An S-dimensional vector, where S is the number of selected eigenvectors (dimensionality reduction), will represent each data point in the transformed subspace. A linear discriminate analysis is helpful for classification and dimensionality reduction problems because it identifies a linear combination of characteristics that maximizes the separation between classes while decreasing the variation within each category.

3.4 prediction of drug sales using enhanced Golden Eagle Optimized Extreme Gradient Boosting (EGEO-Xgboost)

A hybrid AI-based model is developed for medication product sales forecasting using the Enhanced Golden Eagle Optimized and Extreme Gradient Boosting (EGEO-XGBoost) framework. The Enhanced Golden Eagle Optimization (EGO) algorithm provides a global search ability with the strength of predictive accuracy in the Extreme Gradient Boosting (XGBoost) model. This enables model parameter optimization with better precision and efficiency for predicting sales. This helps the model integrate techniques to give pharmaceutical companies the insight needed in managing resources, production planning, and marketing strategy optimization.

3.4.1 Extreme Gradient Boosting (XGBoost)

The XGBoost is a tree-based algorithm that generates trees sequentially to minimize errors, and it further enhances the traditional Gradient Boosting Decision Tree by incorporating regularization to prevent overfitting. It's a highly efficient algorithm and is used widely in different applications. Based on this approach, for drug sales prediction, XGBoost iteratively builds a tree to predict future sales figures and optimizes the model such that overfitting is avoided using a regularization term.

Unlike traditional GBDT, the XGBoost objective function includes a regularization term that balances model accuracy and complexity. The objective function is expressed as Equation (7):

$$U = \sum_{j=1}^{n} S(y_j, \left(F(x_j)\right) + \sum_{h=1}^{t} G\,(f_h) + D \qquad (7)$$

Where, $S(y_j, \left(\ F(x_j)\right)$ represents the loss (error between actual and predicted drug sales), $G(f_h)$ is the regularization term, controlling the complexity of the model, and D is a constant that can be ignored during optimization. The $G(f_h)$ regularization term ensures that over-fitting is also avoided and could be defined by equation (8):

$$G(f_h) = \alpha\mathrm{T} + \frac{1}{2}\eta \sum_{\mathrm{i}=1}^{\mathrm{H}} \mathrm{w}_{\mathrm{i}}^2 \qquad (8)$$

α regulates the penalty for adding leaves to the tree, T is the overall number of trees, and $\mathrm{w}_{\mathrm{i}}^2$ is the weight or predicted sales of each of the leaf nodes. Moreover, the XGBoost optimization uses the second-order Taylor series, in contrast to first-order derivatives with traditional GBDT. Such optimization helps increase the model's capacity to pick up on intricate patterns that are found in data. With MSE used as the loss function, it minimizes error but ensures that no overfitting takes place to obtain accurate predictions. The following is the implementation using the main function equation (9) assuming MSE as a loss function:

$$U = \sum_{j=1}^{n} \left[p_j \omega_{p(y_j)} + \frac{1}{2}(q_j \omega_{q(y_j)}^2)\right] + \alpha T + \frac{1}{2}\eta \sum_{\mathrm{i}=1}^{\mathrm{T}} \omega_i^2 \qquad (9)$$

Where, g_j and T_j represent 1st and 2nd derivatives of loss, which measures the change in the error function concerning the predicted drug sales values. The contribute to the learned parameters of the model for predicting the sales of the drug at each point. $\omega_{p(y_j)}$ and $\omega_{q(y_j)}$ represent the learned parameters of the model for predicting the sales of the drug at each point. The term $q_j \omega_{q(y_j)}^2$ represents the second-order, accounting for curvatures of the loss function, which can promise more stable and accurate predictions for cases involving complicated patterns of drug sales.

The final step calculates the total loss as the summation of losses of all leaf nodes. Since each leaf node is a sale category, the summation of leaf node loss values provides the overall model error. This is described by Equation (10):

$$U = \sum_{i=1}^{H} [p_i \omega_i] + \frac{1}{2}(Q_{\,i} + \eta)\omega_i^2 + \alpha T \qquad (10)$$

Where $p_i = \sum_{j \in I_i} p_j$, $Q_{\,i} = \sum_{j \in I_i} q_j$, which are contributions from each of the leaf nodes that aggregate in their predicted contribution to drug sales.

An optimum prediction of the model is achieved by minimizing this objective function by XGBoost, thus learning the input feature-target output relationship in drug sales. The process includes regularization techniques, which provide good generalizations on new samples by avoiding overfitting problems, thus proving a reliable form of sales forecast.

3.4.2 Enhanced Golden Eagle Optimization (EGEO)

Enhanced Golden Eagle Optimization (EGEO) is the extension of the traditional Golden Eagle Optimization methods. It uses a spiral trajectory mechanism similar to that the golden eagle uses to trace its prey efficiently and strike it down. Thus, EGEO would be able to conduct a more effective global search in the preliminary phases by flying across various regions and a focused local search in the latter phase by honing the attack angle. This makes EGEO equipped with a nice memory component to preserve information about the best solutions it finds so that faster convergence and the least chance of getting trapped into local optima are achieved. With these enhancements, EGEO has a greater degree of robustness and efficiency over traditional methods to solve complex global optimization problems. The mathematical formulae used by golden eagles to simulate their movements while seeking prey were described as follows.

3.4.3 The Spiralling flight patterns of golden eagles

Every AI model within the prediction system remembers the best sales prediction that this model has ever experienced so far. The model, chasing a better drug sales prediction, is simultaneously attracted both to the analysis of history and the prediction of trends. Figure 2 is a graphical representation of the prediction and analysis vectors in 2D space. Every model may randomly pick one sales prediction from the other model and update its parameters based on the data already visited. Also, the model can revisit its memory. Such a process might lead to a sequence of numbers: l ∈ {1, 2 …, N_AI}, where N_AI denotes the total number of AI models.

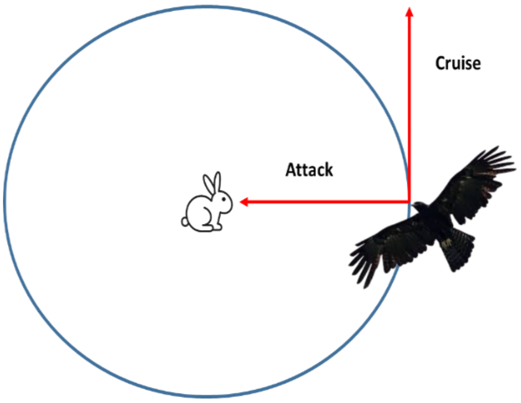

Figure 2: Movement of golden birds in a spiral

3.4.3.1 Selection of Prey

Every AI model should select a target dataset for training and prediction operations in every iteration. The model also chooses its input data from a database of historical drug sales data. Based on this, the selected data computes the prediction and analysis vectors. It checks its memory and updates its memory if the new prediction is better than the earlier one.

3.4.3.2 Attack

A vector that depicts the progress of the prediction from the current state of the model to the best-predicted sales in the model's memory is as follows equation (11):

$$\vec{B}_r = \vec{Y}_l^* - \vec{Y}_r \qquad (11)$$

Where $\vec{B}_r$ is the model's prediction vector, $\vec{Y}_l^*$ is the best previous prediction from the model, and $\vec{Y}_r$ is the current state of the model in the sales prediction process.

3.4.3.3 Travel

The prediction vector runs parallel to the analysis vector, tangent to the predicted sales curve. This is also known as the model's speed of adjustment in sales forecasting. By equation (12) the destination point of the analysis vector is,

$$\vec{C}_r = \frac{d - \Sigma_{f f \neq r} b_f}{b_r} \qquad (12)$$

Whereb_r, $b_f \epsilon \vec{B}_r$ are each component of the hyperplane equation in n-dimensional space, and $\vec{B}_r = \{b_1, b_2, \dots., b_n\}$ represent prediction vector, respectively.

3.4.3.4 Assuming new positions:

The changing direction of its
prediction depends completely on the model's prediction and analysis directions. Consequently, an update step for a model $\boldsymbol{r}$ during an iteration t can be well defined by the following equation:

$$\Delta_{yr} = \vec{j}_1 qb \frac{\vec{B}_r}{\|\vec{B}_r\|} + \vec{j}_2 qc \frac{\vec{C}_r}{\|\vec{C}_r\|} \qquad (13)$$

Where the q_b^t and the q_c^t define at iteration t respectively prediction and analysis coefficients controlling which part of its coefficients have an impact on the prediction or analysis; and the vectors $\vec{j}_1$ and $\vec{j}_2$ are random. The following equation (14) is the new prediction of the model:

$$y_r^{t+1} = y_r^t + \Delta_{yr}^t \qquad (14)$$

When the prediction made at position r is more precise than the preceding position, that new prediction will be updated to memory.

3.4.3.5 Exploration to Exploitation Transition:

The method of the AI model changes from exploring to exploiting. The prediction coefficient q_b and analysis coefficient q_c are used to switch between them. Using equations (15) and (16) are linear expressions:

$$q_b = q_b^0 + \frac{t}{T} |q_b^t - q_b^0| \qquad (15)$$

$$q_c = q_c^0 + \frac{t}{T} |q_c^t - q_c^0| \qquad (16)$$

In the above equation, q_b^0 and q_c^0 are the starting values for predicting the tendency of sales of drugs, q_b, and analyzing market trends, qc, respectively. In this scenario, T denotes the maximum number of iterations, q_b^t, and q_c^t would be the final values for the tendencies to predict and analyze drug sales. The process of the EGEO-XGBoost described in Algorithm 1.

Algorithm:1 The Process of EGEO-XGBoost

```
Step 1: Load the dataset from Kaggle
Step 2: Perform data preprocessing:
  a. Handle missing values
  b. Convert date-time attributes
  c. Normalize numerical features using Min-Max Normalization:
    For each feature X:
    $X_{new} = (X - \min(X)) / (\max(X) - \min(X))$
Step 3: Perform Feature Extraction using Linear Discriminant Analysis (LDA):
  a. Compute mean vector for each class
  b. Compute within-class scatter matrix
  c. Compute between-class scatter matrix
  d. Compute eigenvalues and eigenvectors of $(S_W^{-1} * S_B)$
  e. Select top k eigenvectors and transform the dataset
Step 4: Initialize Enhanced Golden Eagle Optimization (EGEO) algorithm:
  a. Initialize population of golden eagles
  b. Set attack and cruise coefficients
  c. Initialize memory for best solutions
Step 5: Perform iterative optimization:
  a. For each golden eagle:
    i. Select a target prey from memory
    ii. Compute attack vector:
      $B_r = Y_{best} - Y_r$
    iii. Compute cruise vector:
      $C_r = (d - \sum(b_f)) / b_r$
    iv. Update position:
      $\Delta Y_r = j1 * q_b * B_r / ||B_r|| + j2 * q_c * C_r / ||C_r||$
      $Y_r^{t+1} = Y_r^t + \Delta Y_r$
    v. Update memory if new position improves fitness
  b. Adjust attack and cruise coefficients for exploration-exploitation balance:
    $q_b = q_{b0} + (t/T) * |q_{bt} - q_{b0}|$
    $q_c = q_{c0} + (t/T) * |q_{ct} - q_{c0}|$
  c. Repeat until stopping criteria are met
Step 6: Train XGBoost model using optimized hyperparameters:
  a. Define XGBoost objective function:
    $U = \sum(S(y_j, F(x_j))) + \sum(G(f_h)) + D$
  b. Compute regularization term:
    $G(f_h) = \alpha * T + (1/2) * \eta * \sum(w_i^2)$
  c. Use second-order Taylor expansion for loss approximation:
    $U = \sum[p_j * \omega(y_j) + 1/2 * (q_j * \omega_q(y_j)^2)] + \alpha * T + 1/2 * \eta * \sum(w_i^2)$
  d. Optimize quadratic function to minimize loss
  e. Train XGBoost with optimal parameters found via EGEO
Step 7: Evaluate the trained model:
  a. Compute accuracy metrics (RMSE)
  b. Perform validation using test dataset
Step 8: Predict drug sales using the trained EGEO-XGBoost model
Step 9: Output final predictions and insights
```

4 Result

The experiment was performed on a Linux system using an Intel Core 2 Duo processor with 2.16 GHz and 8 GB RAM. The data was retrieved from the Kaggle website and consisted of 600,000 weekly sales records of 57 pharmaceutical products between 2014 and 2019. These included sales date, quantity of drug sold, and the name of the brand. The products were categorized under eight groups using the Anatomical Therapeutic Chemical (ATC) classification, which classifies drugs according to their therapeutic and chemical properties. Figure 3 represents the categories of drugs.

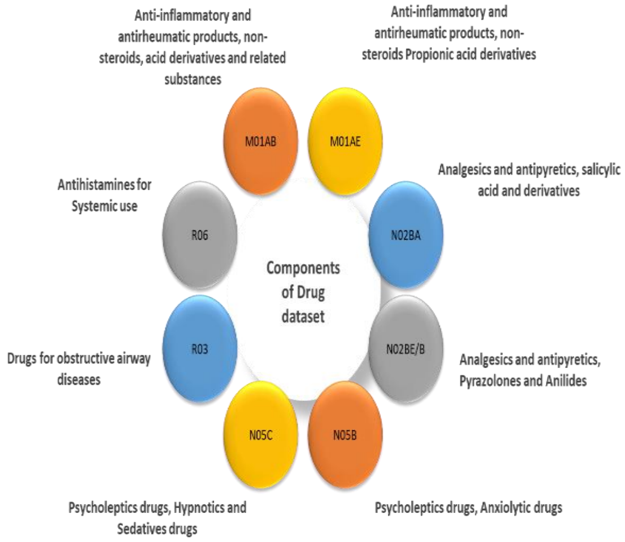
Anti-inflammatory and antirheumatic products, non-steroids, acid derivatives and related substances
Anti-inflammatory and antirheumatic products, non-steroids Propionic acid derivatives
M01AB
M01AE
Antihistamines for Systemic use
Analgesics and antipyretics, salicylic acid and derivatives
R06
N02BA
Components of Drug dataset
R03
N02BE/B
Drugs for obstructive airway diseases
Analgesics and antipyretics, Pyrazolones and Anilides
N05C
N05B
Psycholeptics drugs, Hypnotics and Sedatives drugs
Psycholeptics drugs, Anxiolytic drugs

Figure 3: Elements Of the Drug Dataset

The proposed EGEO-XGBoost model was found to perform better than existing models in most cases with remarkable improvements for M01AB (0.12), M01AE (0.48), N02BA (0.15), N02BE (0.008), N05B (0.006), N05C (0.35), R03 (0.025), and R06 (0.608). This shows that EGEO-XGBoost has better predictive accuracy in drug sales forecasting. Figure 4 illustrates the RMSE comparison for the medicine category using EGEO-XGBoost on the sales prediction dataset.

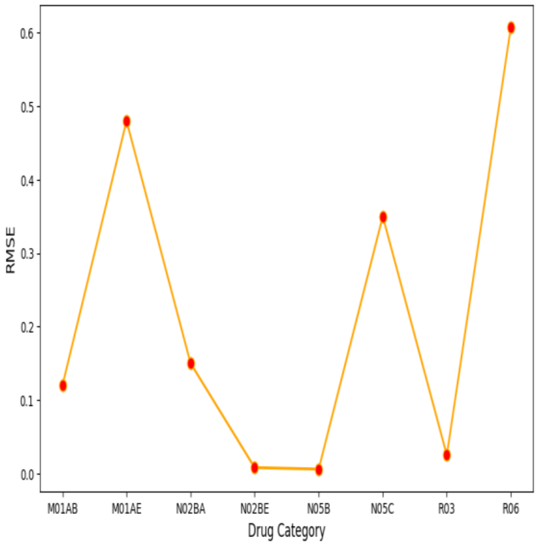

Figure 4: RMSE comparison for drug category using EGEO-XGBoost on sales prediction dataset

For predictive analysis, a few existing models, like RF, MLP, and SVM (Jaganathan et al. [25]), along with Radial Basis Function Neural Network (RBF-NN), Probabilistic NN (P-NN), and Generalized Regression NN (GR-NN) (Rathipriya et al. [24]) were used for analysis. The proposed model, EGEO-XGBoost, was applied using R software to preprocess the data, obtain relevant features from the data, and improves the prediction accuracy.

RMSE: One of the most common metrics used to measure the accuracy of a predictive model. It is defined as the square root of the average of the squared differences between the predicted and observed values. The lower the RMSE, the better the model is performing, as it indicates that the model's predictions are closer to the actual values. Equation (17) was used to calculate the RMSE.

$$RMSE = \sqrt{\frac{1}{2}\sum_{j=1}^{n}(sales\_K_j - sales\_B_j)^2} \quad (17)$$

Table 2 and Figure 5 compare the RMSE for predictive models of drug sales in different categories using existing models (RBF-NN, P-NN, and GR-NN) compared with the proposed EGEO-XGBoost model.

Table 2: RMSE Of Comparison AI-Based model for drug sales prediction

Models	**RMSE**
RBF-NN Rathipriya et al. [24]	6.59
P-NN Rathipriya et al. [24]	6.21
GR-NN Rathipriya et al. [24]	6.01
EGEO-XGBoost [Proposed]	4.02

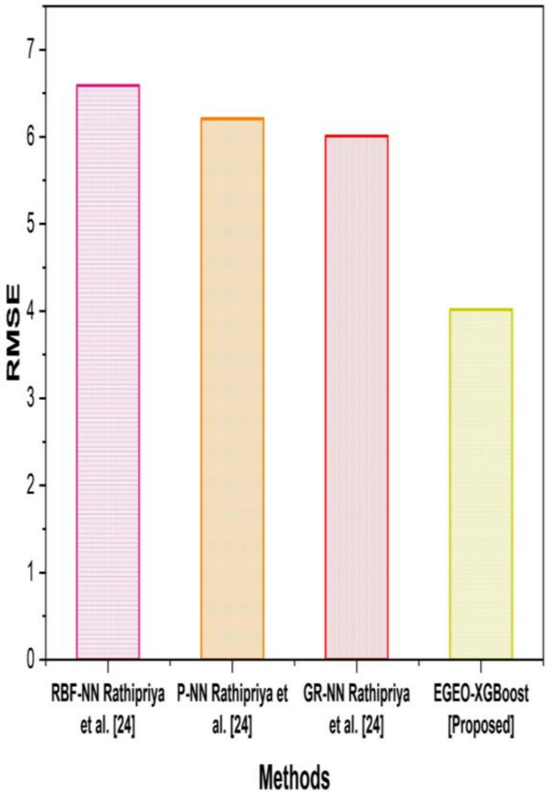

Figure 5: Model performance of RMSE comparison for drug sales prediction

Accuracy: In machine learning, accuracy was a common evaluation metric used to assess a model's performance. Equation (18) was used to calculate accuracy.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \quad (18)$$

Compared to the existing models, such as SVM with an accuracy of 0.82, MLP at 0.79, and RF at 0.77. The proposed EGEO-XGBoost model attains the highest accuracy of 0.90, outperforming all other models. Table 3 and Figure 6 illustrate the outcomes of the accuracy evaluation. The significant improvement indicates the EGEO-XGBoost framework's superior capacity for predictions for drug sales. Figure 6 despite the result of accuracy.

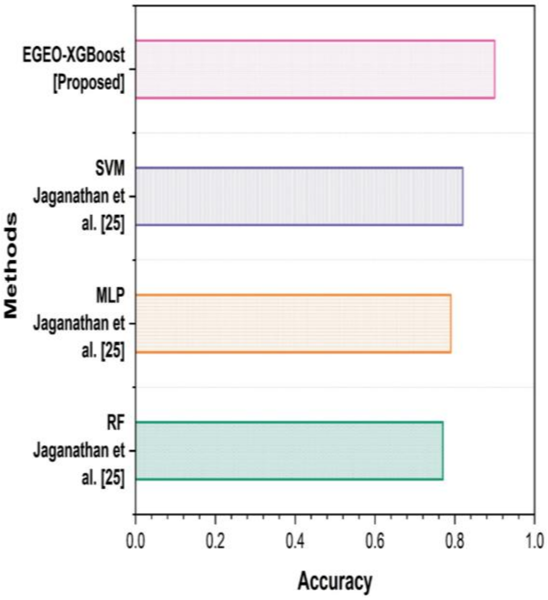

Figure 6: Assessing Overall prediction accuracy in Drug sales forecasting

Sensitivity: Sensitivity measures the model's ability to correctly identify positive instances (accurate predictions of sales growth or success). The calculation of precision is expressed in the following equations (19).

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (19)$$

The EGEO-XGBoost model is sensitivity by 0.92 compared to the other existing approaches: SVM is 0.84, MLP is 0.80, and RF is at 0.81. Sensitivity assessment outcomes are displayed in Table 3 and Figure 7. The higher sensitivity of the EGEO-XGBoost model shows it was able to detect more of the true positives to be important for achieving correct predictions in sales forecasting.

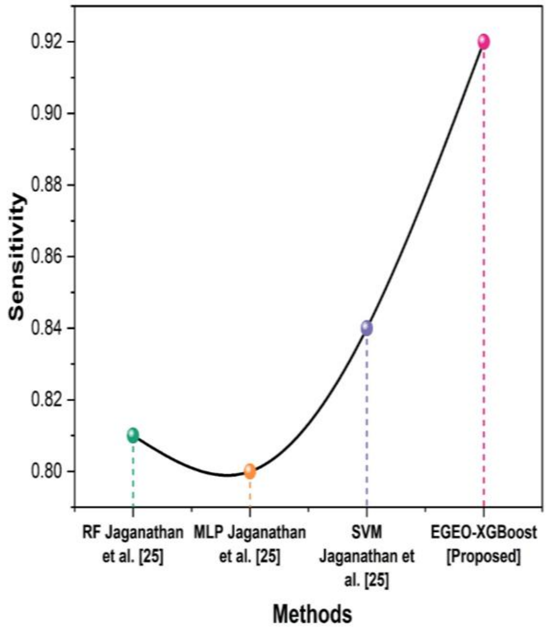

Figure 7: Sensitivity comparison in drug sales prediction

Specificity: Specificity is the measure of how well a model can correctly reject negative instances, meaning that it distinguishes between low sales and other factors that are not relevant. Specificity is calculated using the following equations (20).

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (20)$$

For specificity, the EGEO-XGBoost model also outperformed others with 0.86, which is more than that of SVM, MLP, and RF, having 0.79, 0.78, and 0.69, respectively. The higher specificity of EGEO-XGBoost indicates that this model performs strongly in avoiding false positives, especially in minimizing wrongful sales forecasts of pharmaceutical industries. Figure 8 and Table 3 display the results of the specificity evaluation.

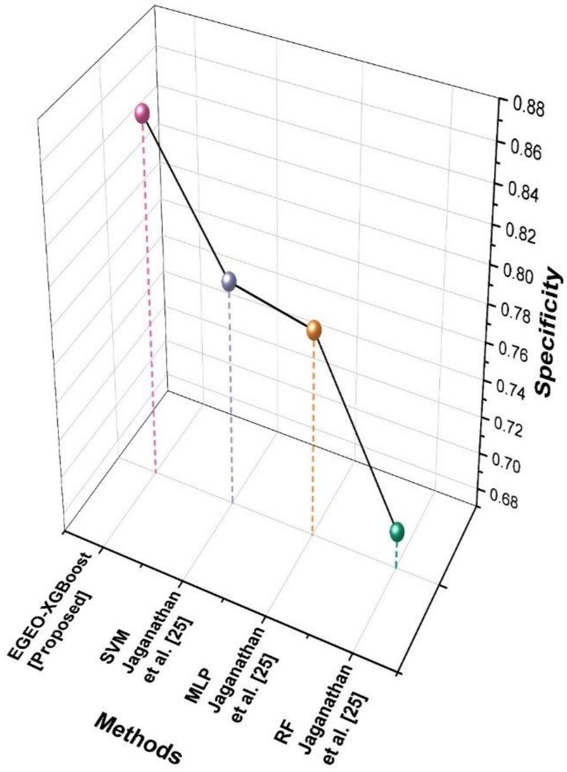

Figure 8: Evaluation specificity in drug sales prediction

Mathew Correlation Coefficient (MCC): MCC is a balanced measure that incorporates all four quadrants of the confusion matrix: True Positives, True Negatives, False Positives, and False Negatives. It ranges between -1 for (total disagreement) up to +1 for (perfect agreement), with a value of 0 indicating no better than random prediction.

$$MCC = \frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (21)$$

The MCC score of the EGEO-XGBoost model is 0.82, which is significantly higher than that of SVM (0.61), MLP (0.59), and RF (0.57). This higher MCC indicates that the EGEO-XGBoost model has better overall classification performance, meaning a better balance between true positives and negatives. Table 3 and Figure 9 show the MCC evaluation results.

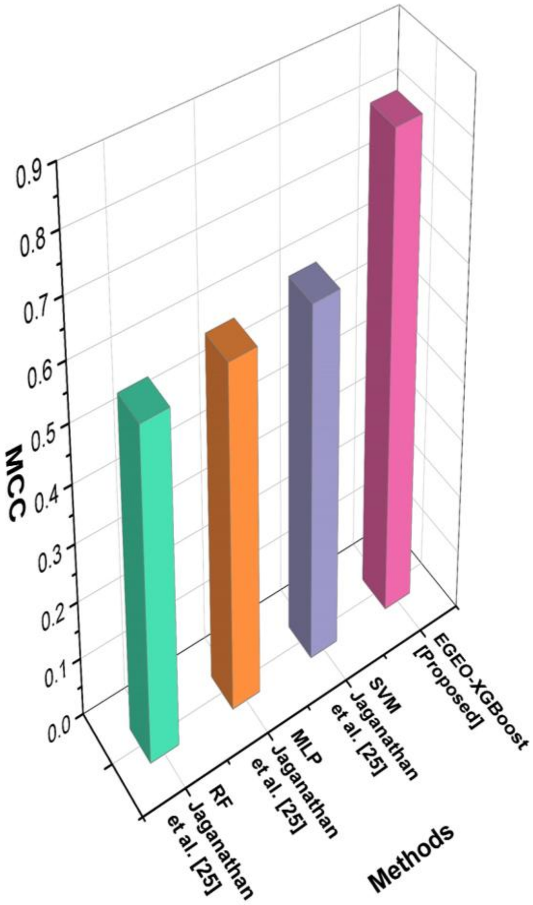

Figure 9: Evaluation of MCC comparison in drug sales forecasting

F1-score: It is the harmonic mean of precision and sensitivity that balances the two. It is useful when both false positives and false negatives matter for a very imbalanced dataset. Equation 18, therefore, was used to calculate the F1 score.

$$F1 - Score = \frac{2TP}{2TP+FP+FN} \quad (22)$$

The EGEO-XGBoost model performs well with an F1-score of 0.94, which is much higher than other existing methods SVM (0.84), MLP (0.80), and RF (0.77). This indicates that a higher F1-score shows the EGEO-XGBoost to achieve an optimal balance between precision and recall in terms of its performance for the balanced prediction of drug sales. Figure 10 and Table 3 demonstrate the evaluation result of the F1score.

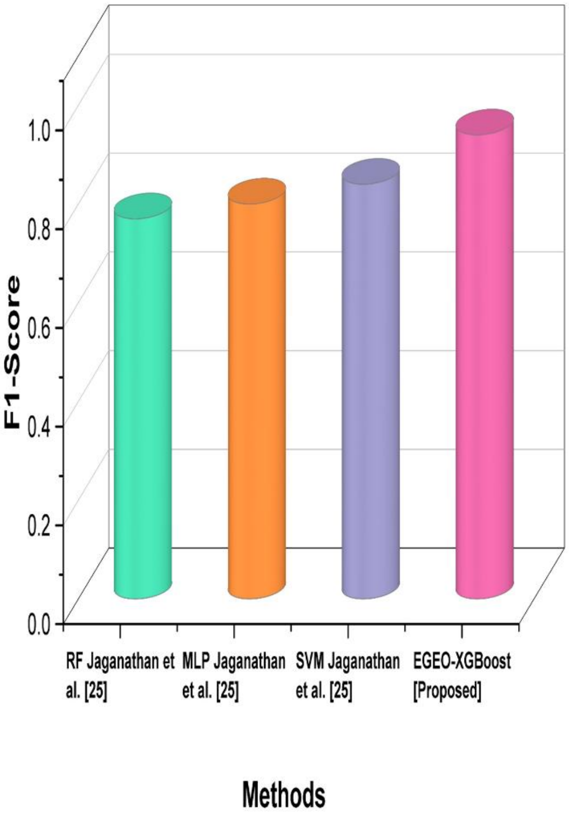

Figure 10: F1-score performance in drug sales prediction

Table 3: Performance comparison of Different models for drug sales prediction

Models	Accuracy	Sensitivity	Specificity	MCC	F1score
RF Jaganathan et al. [25]	0.77	0.81	0.69	0.57	0.77
MLP Jaganathan et al. [25]	0.79	0.80	0.78	0.59	0.80
SVM Jaganathan et al. [25]	0.82	0.84	0.79	0.61	0.84
EGEO-XGBoost [Proposed]	0.90	0.92	0.86	0.82	0.94

5 Discussion

Evaluation shows a significant increase in the accuracy of forecasting over existing models, namely, RBF-NN, P-NN, GR-NN, SVM, MLP, and RF, concerning the development and implementation of the EGEO-XGBoost model for predicting drug sales. The model also achieved (0.90) accuracy, which is more than SVM at 0.82, MLP at 0.79, and RF at 0.77. As for sensitivity, EGEO-XGBoost was at 0.92, more than SVM at 0.84, MLP at 0.80, and RF at 0.81, indicating it is better at correctly predicting positive sales instances. For instance, the score of EGEO-XGBoost was impressive at (0.86), which was higher than SVM (0.79), MLP (0.78), and RF (0.69), meaning it is less likely to give false positives. The MCC score of 0.82 for EGEO-XGBoost also indicates that it has a better balance between true positives and true negatives compared to SVM (0.61). Last, the model attained an F1-score of (0.94), much higher than other models, meaning that it balances precision and recall very well. EGEO-XGBoost outperforms other models in all key evaluation metrics and, therefore, can be regarded as a highly effective tool for drug sales prediction.

6 Conclusion

The EGEO-XGBoost model has highly effective performance regarding the prediction of drug sales, with higher predictive accuracy values compared to any traditional model. Using advanced techniques in the form of Min-max normalization, application of Linear Discriminant Analysis in feature extraction, and the Enhanced Golden Eagle Optimization, the proposed model shows significant enhancement in key metrics such as accuracy (0.90), sensitivity (0.92), specificity (0.86), MCC (0.82), F1-score (0.94) and RMSE (4.02). These findings thus underscore the ability of AI-driven models to enhance drug sales forecasting, leading to better decision-making, resource management, and marketing strategies in the pharmaceutical industry. Incorporation of such models would help in improved operational efficiency and better planning in the pharmaceutical industry.

Limitation and future scope:

Although promising results are produced, the limitations of the model lie in reliance on historical data that may not reflect sudden changes in the market or external influences. Moreover, the EGEO-XGBoost model needs further optimization for large datasets originating from heterogeneous pharmaceutical markets. The future direction can be oriented towards real-time data integration and the development of hybrid models, making it even more adaptable and robust for use in global drug sales forecasting.

References

[1] Jakovljevic, M., Liu, Y., Cerda, A., Simonyan, M., Correia, T., Mariita, R.M., Kumara, A.S., Garcia, L., Krstic, K., Osabohien, R., and Toan, T.K., The Global South political economy of health financing and spending landscape–history and presence, Journal of Medical Economics, vol. 24, sup1, pp. 25-33, 2021. https://doi.org/10.1080/13696998.2021.1915807.

[2] Ren, S., Chan, H.L., and Siqin, T., Demand forecasting in retail operations for fashionable products: methods, practices, and real case study, Annals of Operations Research, vol. 291, pp. 761-777, 2020. https://doi.org/10.1007/s10479-019-03479-w

[3] Feizabadi, J., Machine learning demand forecasting and supply chain performance, International Journal of Logistics Research and Applications, vol. 25, no. 2, pp. 119-142, 2022. https://doi.org/10.1080/13675567.2021.1898171

[4] Kumar, S.A., Ananda Kumar, T.D., Beeraka, N.M., Pujar, G.V., Singh, M., Narayana Akshatha, H.S., and Bhagyalalitha, M., Machine learning and deep learning in data-driven decision-making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry, Future Medicinal Chemistry, vol. 14, no. 4, pp. 245-270, 2022. https://doi.org/10.4155/fmc-2021-0215

[5] Roy, S.N., Mishra, S., and Yusof, S.M., The emergence of drug discovery in machine learning, Technical Advancements of Machine Learning in Healthcare, pp. 119-138, 2021. https://doi.org/10.1201/9781003116886-10

[6] Sugandha, S., Choubey, R.R., Gupta, R.K., and Gupta, S.B., Role of digital transformation and technology adoption in the efficiency of the pharmaceutical industry.

[7] Meenakshi, D.U., Nandakumar, S., Francis, A.P., Sweety, P., Fuloria, S., Fuloria, N.K., Subramaniyan, V., and Khan, S.A., Deep learning and site-specific drug delivery: The future and intelligent decision support for pharmaceutical manufacturing science, Deep Learning for Targeted Treatments: Transformation in Healthcare, pp. 1-38, 2022. https://doi.org/10.1002/9781119879816.ch1

[8] Baviskar, K., Bedse, A., Raut, S., and Darapaneni, N., Artificial Intelligence and Machine Learning-Based Manufacturing and Drug Product Marketing, Bioinformatics Tools for Pharmaceutical Drug Product Development, pp. 197-231, 2023. https://doi.org/10.1002/9781119859115.ch8

[9] Srivastava, D., Soni, D., Sharma, V., Kumar, P., and Singh, A.K., An artificial intelligence-based recommender system to analyze drug target indication for drug repurposing using linear machine learning

algorithm, Journal of Algebraic Statistics, vol. 13, no. 3, pp. 790-797, 2022.
https://doi.org/10.1016/j.jastat.2021.12.001

[10] Adepu, A., and Poonia, P., Artificial intelligence-based systems direct the drugs to the faulty DNA sequences and tailor the gene sequences, Artificial Intelligence, vol. 52, no. 4, 2023.

[11] Bhattamisra, S.K., Banerjee, P., Gupta, P., Mayuren, J., Patra, S., and Candasamy, M., Artificial intelligence in pharmaceutical and healthcare research, Big Data and Cognitive Computing, vol. 7, no. 1, p. 10, 2023.
https://doi.org/10.3390/bdcc7010010

[12] Vora, L.K., Gholap, A.D., Jetha, K., Thakur, R.R.S., Solanki, H.K., and Chavda, V.P., Artificial intelligence in pharmaceutical technology and drug delivery design, Pharmaceutics, vol. 15, no. 7, p. 1916, 2023.
https://doi.org/10.3390/pharmaceutics15071916

[13] Shilong, Z., The machine learning model for sales forecasting by using XGBoost, in 2021 IEEE International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 480-483, IEEE, 2021.
https://doi.org/10.1109/ICCECE51280.2021.9342324

[14] Lin, S., Wang, Y., Zhang, L., Chu, Y., Liu, Y., Fang, Y., Jiang, M., Wang, Q., Zhao, B., Xiong, Y., and Wei, D.Q., MDF-SA-DDI: Predicting drug–drug interaction events based on multi-source drug fusion, multi-source feature fusion, and transformer self-attention mechanism, Briefings in Bioinformatics, vol. 23, no. 1, p. bbab421, 2022.
https://doi.org/10.1093/bib/bbab421

[15] Tichy, E.M., Schumock, G.T., Hoffman, J.M., Suda, K.J., Rim, M.H., Tadrous, M., Stubbings, J., Cuellar, S., Clark, J.S., Wiest, M.D., and Matusiak, L.M., National trends in prescription drug expenditures and projections for 2020, American Journal of Health-System Pharmacy, vol. 77, no. 15, pp. 1213-1230, 2020.
https://doi.org/10.1093/ajhp/zxaa186

[16] Agrawal, G., Tushir, S., Arora, D., and Sangwan, K., Artificial intelligence in pharmaceutical drug delivery, in 2024 International Conference on Computational Intelligence and Computing Applications (ICCICA), vol. 1, pp. 406-410, IEEE, 2024.

[17] Biehn, S.E., Goncalves, L.M., Lehmann, J., Marty, J.D., Mueller, C., Ramirez, S.A., Tillier, F., and Sage, C.R., BioPrint meets the AI age: Development of artificial intelligence-based ADMET models for the drug-discovery platform SAFIRE, Future Medicinal Chemistry, vol. 16, no. 7, pp. 587-599, 2024.
https://doi.org/10.4155/fmc-2024-0007

[18] Saikia, S., Prajapati, J.B., Prajapati, B.G., Padma, V.V., and Pathak, Y.V., The role of artificial intelligence in therapeutic drug monitoring and clinical toxicity, in Recent Advances in Therapeutic Drug Monitoring and Clinical Toxicology, pp. 67-85, Springer International Publishing, 2022.
https://doi.org/10.1007/978-3-031-12398-6_5

[19] Ali, A.M.A., and Alrobaian, M.M., Strengths and weaknesses of current and future prospects of artificial intelligence-mounted technologies applied in the development of pharmaceutical products and services, Saudi Pharmaceutical Journal, 2024, p. 102043.
https://doi.org/10.1016/j.jsps.2024.102043

[20] Łapińska, N., Pacławski, A., Szlek, J., and Mendyk, A., SerotoninAI: Serotonergic system-focused artificial intelligence-based application for drug discovery, Journal of Chemical Information and Modeling, vol. 64, no. 7, pp. 2150-2157, 2024.
https://doi.org/10.1021/acs.jcim.3c01517

[21] Serrano, D.R., Luciano, F.C., Anaya, B.J., Ongoren, B., Kara, A., Molina, G., Ramirez, B.I., Sánchez-Guirales, S.A., Simon, J.A., Tomietto, G., and Rapti, C., Artificial intelligence (AI) applications in drug discovery and drug delivery: Revolutionizing personalized medicine, Pharmaceutics, vol. 16, no. 10, p. 1328, 2024.
https://doi.org/10.3390/pharmaceutics16101328

[22] Jena, G.K., Patra, C.N., Jammula, S., Rana, R., and Chand, S., Artificial intelligence and machine learning-implemented drug delivery systems: A paradigm shift in the pharmaceutical industry, Journal of Bio-X Research, vol. 7, p. 0016, 2024.
https://doi.org/10.34133/jbioxresearch.0016

[23] Visan, A.I., and Negut, I., Integrating artificial intelligence for drug discovery in the context of revolutionizing drug delivery, Life, vol. 14, no. 2, p. 233, 2024.
https://doi.org/10.3390/life14020233

[24] Rathipriya, R., Abdul Rahman, A.A., Dhamodharavadhani, S., Meero, A., and Yoganandan, G., Demand forecasting model for time-series pharmaceutical data using shallow and deep neural network models, Neural Computing and Applications, vol. 35, no. 2, pp. 1945-1957, 2023.
https://doi.org/10.1007/s00521-022-07889-9

[25] Jaganathan, K., Tayara, H., and Chong, K.T., Prediction of drug-induced liver toxicity using SVM and optimal descriptor sets, International Journal of Molecular Sciences, vol. 22, no. 15, p. 8073, 2021.
https://doi.org/10.3390/ijms22158073