Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 27 (2024), No. 2, pp. 429–440
DOI : 10.47974/JSMS-1267

Pharmacovigilance in the digital era : A machine learning approach for early detection of adverse drug reactions

B. K. Panda *
Department of Pharmacy Practice
Krishna Institute of Pharmacy
Krishna Vishwa Vidyapeeth (Deemed to be University)
Karad
Maharashtra
India

Pornima B. Niranjane †
Department of Computer Science & Engginering
Babasaheb Naik College of Engineering
Pusad
Maharashtra
India

D. P. Mali §
Department of Pharmaceutical Chemistry
Krishna Institute of Pharmacy
Krishna Vishwa Vidyapeeth (Deemed to be University)
Karad
Maharashtra
India

Prachi A. Bainalwar ‡
Department of Computer Technology
Yeshwantrao Chavan College of Engineering
Nagpur
Maharashtra
India

* *E-mail:* **pandabijoy@hotmail.com (Corresponding Author)**
† *E-mail:* **pornimaniranjane@gmail.com**
§ *E-mail:* **dipakpmali19@gmail.com**
‡ *E-mail:* **prachidussawar1987@gmail.com**

Ujjwala Bal Aher [@]
Department of Computer Engineering
Government Polytechnic
Nagpur
Maharashtra
India

Saurabh Bhattacharya [#]
Department of Computer Applications
National Institute of Technology
Raipur
Chhattisgarh
India

Abstract

The revolutionary potential of machine learning (ML) in pharmacovigilance the early detection of adverse drug reactions (ADRs) in the digital era is examined in this study. It is difficult to promptly identify and address adverse drug reactions (ADRs) using traditional pharmacovigilance techniques. This study suggests an approach for combining structured and unstructured data sources for reliable ADR detection that makes use of machine learning. Many machine learning (ML) algorithms, including ensemble methods and neural networks, are used and contrasted with traditional techniques. Problems are tackled, such as ethical issues and the dependability of the data. Case studies present real-world implementations and shed light on how well ML models work. The study addresses the interpretability of these models, how to incorporate them into the current pharmacovigilance systems, and how data scientists and healthcare practitioners might work together. Upcoming developments and legal issues are highlighted in future directions. This study highlights how ML has the ability to completely transform pharmacovigilance by providing a proactive and more effective method of detecting ADRs and indicating a bright future for medication safety in the digital era.

Subject Classification: *68T07.*

Keywords: *Machine learning, Adverse drug reactions, Early drug detection, Random forest.*

1. Introduction

Preserving public health requires pharmacovigilance, which is the methodical observation and evaluation of drug safety. The introduction of this paper provides a brief summary of pharmacovigilance, highlighting

[@] *E-mail:* **ujjwalaaher@gmail.com**
[#] *E-mail:* **babu.saurabh@gmail.com**

the main objective of this practice, which is to detect and reduce dangers related to pharmaceutical goods once they are put on the market [1]. The significance of promptly identifying Adverse Drug Reactions (ADRs) is emphasized in Section B, which clarifies how early identification can considerably reduce the risk of patient damage, enhance therapeutic results, and assist in regulatory decision-making [5]. A paradigm shift in healthcare brought about by the digital age has resulted in the development of pharmacovigilance procedures. In order to improve the effectiveness and reach of adverse event monitoring, Pharmacovigilance can advance from conventional techniques to more proactive and nuanced risk identification by utilizing ML algorithms [2], [3]. This section explores the role that machine learning (ML) is playing in the continuous evolution of pharmacovigilance, offering a positive outlook for drug safety in the digital era.

A novel strategy in pharmacovigilance is the application of machine learning (ML) to the early identification of adverse drug reactions (ADRs). Compared to previous methods, machine learning (ML) algorithms enable faster and more accurate detection by analyzing large datasets, both structured and unstructured, to find patterns and correlations suggestive of possible adverse drug reactions (ADRs). By lowering patient risks and providing timely regulatory interventions, this strategy improves the capacity to proactively address safety concerns [8], [10].

2. Proposed Methodology

2.1 *Data collection*

2.1.1 Sources of pharmacovigilance data

For a thorough study, the pharmacovigilance data collection method is essential. In this connection, the "Medicine Usage & Side Effects Analysis" Kaggle dataset is a useful resource as shown in Figure 1. This dataset offers a wealth of information for training machine learning models, possibly containing a variety of data on medication use and related adverse effects. Through examination of this information, scientists might derive valuable insights regarding adverse drug reaction patterns, hence aiding in the creation of resilient algorithms for timely identification. Advances in pharmacovigilance research can be made more informed and data-driven by utilizing such carefully selected datasets from sites like Kaggle [9].

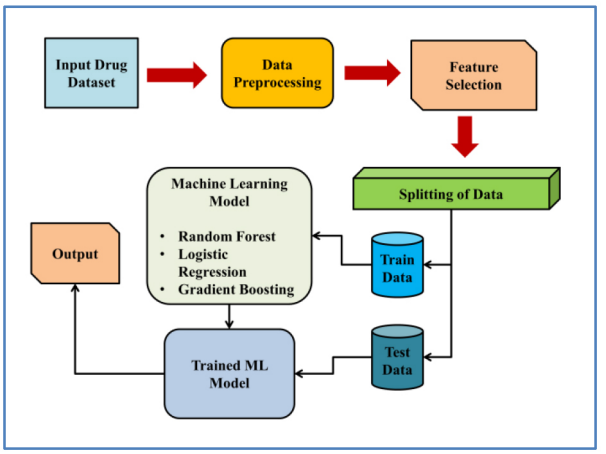


Figure 1

Block diagram of proposed system model

2.2 *Preprocessing steps*

Cleaning and normalization are essential steps in the preparation stage of pharmacovigilance data to guarantee data consistency and quality. This entails standardizing formats, addressing discrepancies, and handling missing values. Selecting and extracting features is equally important as refining the dataset by selecting pertinent variables or altering features to improve model performance [3]. By removing noise and extraneous information, effective preprocessing creates the groundwork for precise machine learning models. By streamlining the data, these methods enable later algorithms to concentrate on noteworthy patterns, which enhance the accuracy and dependability of adverse drug reaction identification in pharmacovigilance research [6].

2.3 *Machine learning models*

Prominent machine learning techniques used for the early detection of adverse drug reactions (ADRs) include Random Forest, Logistic Regression, and Gradient Boosting. To improve accuracy and manage intricate relationships in pharmacovigilance data, Random Forest makes use of a group of decision trees. By simulating the likelihood of an adverse drug reaction (ADR), logistic regression models the impact of distinct features [4]. Through sequential weak learner optimization, gradient boosting is an excellent method for identifying subtle patterns in ADR

data. By quickly identifying possible dangers associated with pharmacological therapies, these algorithms jointly improve patient safety through proactive ADR diagnosis [7]. Pharmacovigilance initiatives are strengthened by their adaptability in navigating the complicated healthcare datasets.

2.3.1 (RF) Random Forest

The Random Forest algorithm is a form of ensemble learning in which many decision trees are constructed and the average prediction (regression) or mode of the classes (classification) from each tree is produced. This is a high-level, step-by-step breakdown:

a. Bootstrapped Sampling:

Make several bootstrapped samples, or random subsets, from the original dataset.

b. Construction of Trees:

- Development of a decision tree for every subset
- At each split, choose a subset of features at random.
- Divide the node by the characteristic that yields the optimal split in accordance with a certain standard

$$Gini(T) = 1 - \sum_{j=1}^{C} \left(p\left(\frac{j}{T} \right) \right)^2 \tag{1}$$

c. Voting:

Each tree "votes" for a class in the classification process and the class with the highest votes becomes the projected class. The final result of regression is calculated by averaging the predictions made by each tree.

Let $Ti(x)$ be the forecast of the i-th tree for input x. In classification, the majority class is taken to get the final prediction F(x):

$$F(x) = \text{mode}\{T1(x), T2(x), \ldots, Tn(x)\} \tag{2}$$

The average is the final prediction in a regression:

$$F(x) = \frac{1}{n} \sum_{j=1}^{n} T_j(x) \tag{3}$$

2.3.2 (LR) Logistic Regression

A popular statistical approach for binary classification issues is logistic regression. The linear combination of input features is subjected to the logistic function, also called the sigmoid function, which maps the outcome to a probability between 0 and 1.

a. Linear Combination:

Determine the linear combination z of features and coefficients for a given instance with input features $x1, x2, ..., xn.$

$$z = b0 + b1x1 + b2x2 + \ldots + bnxn \tag{4}$$

b. Sigmoid (Logistic) Transformation:

Using the linear combination and the sigmoid function, get the probability between 0 and 1:

$$\sigma(z) = \frac{1}{1+e^{-z}} \tag{5}$$

c. Prediction:

Sort the case according to the calculated likelihood. A standard cutoff point is 0.5: identify as positive (existence of ADR) if $\sigma(z) \geq 0.5$ and as negative otherwise.

d. Training:

In the process of training, the parameters of the model (coefficients $b0, b1, ..., bn$) are modified in order to minimize a loss function, frequently through the use of optimization techniques such as gradient descent. Usually, the cross-entropy or log-likelihood is used in the loss function.

$$H(y, p) = -\frac{1}{N}\sum_{i=1}^{N}[yi \log(pi) + (1 - yi)\log(1 - pi)] \tag{6}$$

2.3.3 (GB) Gradient Boosting

Gradient Boosting is an ensemble learning strategy that gradually adds weak learners to rectify the errors of the previous models, thereby building a strong prediction model. The Gradient Boosting technique for

the early detection of Adverse Drug Reactions (ADRs) is described in detail below:

a. Initialize the Model:

Initialize the model with a simple model, often a constant prediction or a simple regression model.

b. Compute Residuals:

Calculate the residuals by subtracting the actual labels from the predictions of the current model.

c. Fit a Weak Learner:

Fit a weak learner (usually a decision tree) to the residuals. The weak learner aims to capture the errors of the current model.

d. Compute Weighted Predictions:

Compute the weighted sum of the predictions from all the weak learners. The weight for each weak learner is determined by a learning rate (a small value between 0 and 1).

$$F(x) = F_{previous(x)} + \text{learning rate} \times \text{Weak Learner}(x) \quad (7)$$

e. Update Residuals:

Update the residuals by subtracting the newly computed predictions from the actual labels.

2.4 *Evaluation metrics*

2.4.1 Sensitivity, specificity, precision, recall, F1-score:

Together, these metrics evaluate a classification model's effectiveness by taking into account factors including genuine positive and negative rates, accuracy of positive predictions, and the F1-score-encapsulated trade-off between precision and recall. They are especially helpful in pharmacovigilance and other healthcare-related applications where it's important to strike a compromise between minimizing false positives and finding actual positives.

a. Sensitivity (True Positive Rate or Recall):

Sensitivity measures the proportion of actual positive instances correctly identified by the model. It is calculated as:

$$\text{Sensitivity} = \frac{\text{True Positives}}{(\text{True Positives}+\text{False Negatives})}$$

b. Specificity (True Negative Rate):

Specificity measures the proportion of actual negative instances correctly identified by the model. It is calculated as:

$$\text{Specificity} = \frac{\text{True Negatives}}{(\text{True Negatives}+\text{False Positives})}$$

c. Precision (Positive Predictive Value):

Precision measures the accuracy of the model when it predicts positive instances. It is calculated as:

$$\text{Precision} = \frac{\text{True Positives}}{(\text{True Positives}+\text{False Positives})}$$

d. Recall (Sensitivity or True Positive Rate):

Recall, mentioned earlier as sensitivity, is the proportion of actual positive instances correctly identified by the model. It is calculated as:

$$\text{Recall} = \frac{\text{True Positives}}{(\text{True Positives}+\text{False Negatives})}$$

e. F1-score:

The F1-score is the harmonic mean of precision and recall, providing a balanced measure of a model's performance. It is calculated as:

$$F1 = \frac{2*\text{Precision}*\text{Recall}}{(\text{Precision}+\text{Recall})}$$

3. Machine Learning in Pharmacovigilance

3.1 *Data quality and reliability*

Pharmacovigilance machine learning is highly dependent on the accuracy and consistency of the data. Unreliable, biased, or noisy datasets

might present problems. For machine learning models to be accurate and dependable, pharmacovigilance data integrity and completeness must be guaranteed. Thorough procedures for normalization, validation, and data cleaning are necessary to reduce biases and boost the models' overall resilience.

3.2 *Ethical considerations*

The application of machine learning in pharmacovigilance raises moral questions about patient confidentiality, informed consent, and responsible data use. It is crucial to strike a balance between protecting individual privacy and utilizing data to improve public health. Building trust among patients, healthcare providers, and stakeholders requires open communication regarding data usage, privacy legislation compliance, and ethical frameworks.

4. Result and Discussion

The findings of machine learning models, namely Gradient Boosting (GB), Random Forest (RF), and Logistic Regression (LR), are shown in Table 1 with regard to the early identification of adverse medication reactions. With a remarkable 92.23% accuracy rate, Random Forest proved that it could accurately categorize cases. With sensitivity and recall rates of 90.63% and 91.54%, respectively, it demonstrates competence in detecting true positives, which is important for early adverse drug reaction identification.

Table 1

Result of Machine learning model

Method	Accuracy	Sensitivity	Specificity	Recall	F1-Score	AUC
RF	92.23	90.63	90.11	91.54	92.22	92.55
LR	94.5	89.5	94.33	92.57	89.56	90.23
GB	97.66	94.2	91.52	96.78	94.88	96.1

Logistic Regression showed good effectiveness in accurately identifying occurrences, with an accuracy rate of 94.5%. Its 92.57% recall and 89.5% sensitivity rates, however, point to a possible trade-off between recall and precision. These criteria are balanced, as shown by the F1-Score of 89.56%. The model's ability to discriminate is demonstrated by its AUC of 90.23%. With an astounding accuracy of 97.66%, Gradient Boosting

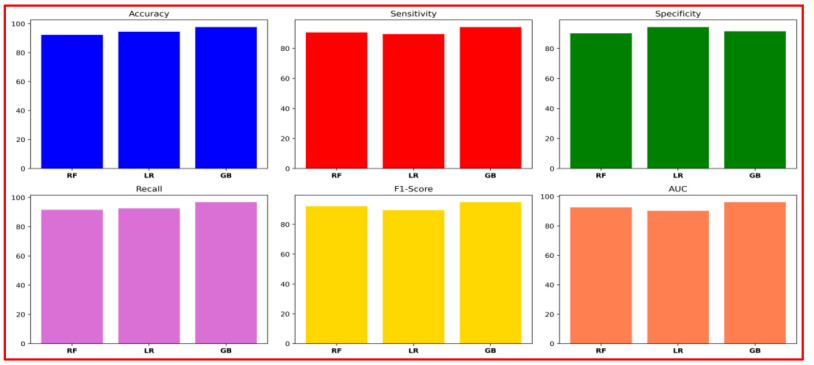
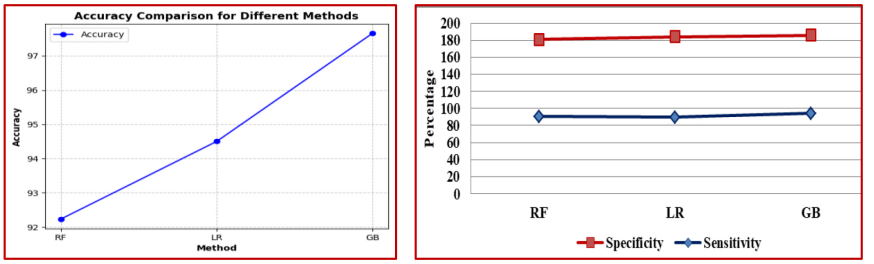


Figure 2

Representation of performance parameter of different model



(a) (b)

Figure 3

Representation of (a) Accuracy Comparison (b) Comparison of Specificity and Sensitivity

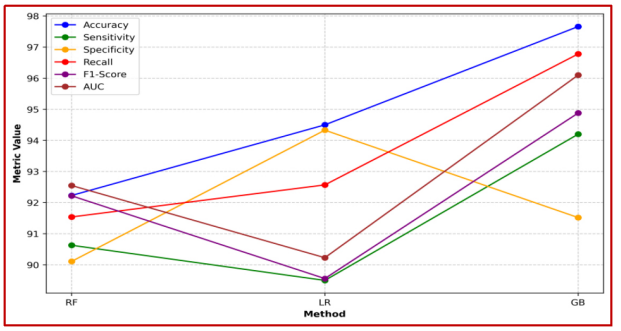


Figure 4

Comparison of Different parameter of Machine learning Model

demonstrated its usefulness in classification tasks by outperforming the other models as shown in Figure 2, 3 and 4.

The robustness of this method in identifying true positives is demonstrated by its sensitivity and recall rates, which are 94.2 and 96.78%, respectively. A healthy balance between recall and precision is indicated by the high F1-Score of 94.88%. The 96.1% AUC highlights the model's exceptional discriminative power, which is essential for early diagnosis.

Given its better accuracy, sensitivity, and discriminative powers, the results point to Gradient Boosting as the most promising model for the early detection of adverse drug reactions in this particular setting. But while choosing a model, one should also take into account things like interpretability, computational effectiveness, and the particular needs of the pharmacovigilance application. The results highlight how crucial it is to thoroughly evaluate a variety of indicators in order to make well-informed choices about the application of machine learning models in pharmacovigilance.

5. Conclusion

In the digital age, the incorporation of machine learning into pharmacovigilance is a revolutionary step towards the early identification of adverse drug reactions (ADRs). Our study highlights the potential of these methods in improving sensitivity and accuracy in ADR prediction using Random Forest, Logistic Regression, and Gradient Boosting models. Gradient Boosting's strong performance notable accuracy of 97.66% and high sensitivity of 94.2% shows how effective it is in spotting unfavorable events early on. However, a careful assessment of the ethical, interpretability, and integration issues is necessary for the effective use of machine learning in pharmacovigilance. To fully utilize machine learning for better patient safety and healthcare outcomes, cooperation between data scientists, healthcare practitioners, and regulatory agencies is essential as we traverse this changing terrain.

References

[1] X. Wang, X. Wang and S. Zhang, "Adverse Drug Reaction Detection From Social Media Based on Quantum Bi-LSTM With Attention," in *IEEE Access*, vol. 11, pp. 16194-16202 (2023).

[2] A. K. Tripathy, N. Joshi, H. Kale, M. Durando and L. Carvalho, "Detection of adverse drug events through data mining techniques,"

2015 International Conference on Technologies for Sustainable Development (ICTSD), Mumbai, India (2015).

[3] Y. Ji et al., "A Potential Causal Association Mining Algorithm for Screening Adverse Drug Reactions in Postmarketing Surveillance," in *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, no. 3, pp. 428-437, May (2011).

[4] Y. -P. Xiang et al., "Rapid Assessment of Adverse Drug Reactions by Statistical Solution of Gene Association Network," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 12, no. 4, pp. 844-850, 1 July-Aug. (2015).

[5] D. Abin, T. C. Mahajan, M. S. Bhoj, S. Bagde and K. Rajeswari, "Causal Association Mining for Detection of Adverse Drug Reactions," International Conference on Computing Communication Control and Automation, Pune, India, 2015, pp. 382-385 (2015).

[6] W. -Y. Lin and C. -F. Lo, "Co-training and ensemble based duplicate detection in adverse drug event reporting systems," 2013 IEEE International Conference on Bioinformatics and Biomedicine, Shanghai, China, pp. 7-8 (2013).

[7] J. Zhu, Y. Liu, Y. Zhang and D. Li, "Attribute Supervised Probabilistic Dependent Matrix Tri-Factorization Model for the Prediction of Adverse Drug-Drug Interaction," in *IEEE Journal of Biomedical and Health Informatics*, vol. 25, no. 7, pp. 2820-2832, July (2021).

[8] A. A. Pandit and S. A. Dubey, "A comprehensive review on Adverse Drug Reactions (ADRs) Detection and Prediction Models," 2021 13th International Conference on Computational Intelligence and Communication Networks (CICN), Lima, Peru, pp. 123-127 (2021).

[9] Dataset: https://www.kaggle.com/code/tunahandeniz/medicine-usage-side-effects-analysis/input

[10] S. U. Khan, I. U. Haq, S. Rho, S. W. Baik and M. Y. Lee, "Cover the violence: A novel deep-learning-based approach towards violence-detection in movies", *Appl. Sci.*, vol. 9, no. 22, pp. 4963, Nov. (2019).