International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

AI-Driven Drug Discovery Platform

G.Swapna[1],E.Susmitha[2],G.Kavyamrutha[3],J.Hyma4,M.Sameera[5]
[1,2]Assistant Professors, Department Of Computer Science and Engineering, Rajiv Gandhi University Of Knowledge Technologies RK Valley, Kadapa, 516330, Andhra Pradesh, India.
[3,4,5]Students, Department Of Computer Science And Engineering,Rajiv Gandhi University Of Knowledge Technologies RK Valley, Kadapa, 516330, Andhra Pradesh, India.
Email ID: *gswapna51@gmail.com[1] , susmisuni@gmail.com[2] , kavyamrutha.g@gmail.com[3] , jinkalahyma2005@gmail.com[4] , msameera2610@gmail.com[5]*

Abstract
The process of drug discovery using traditional methods is time-consuming, expensive, and often inefficient in meeting urgent medical needs. This study presents an AI-driven drug discovery platform designed to assist healthcare professionals in identifying effective treatments quickly and accurately. The primary aim of this research is to reduce the time required for drug identification and improve decision-making in medical applications. The proposed system utilizes Artificial Intelligence techniques to analyze disease inputs and recommend the top three suitable drugs based on key parameters such as absorption and toxicity. In addition, the platform provides insights into possible deficiencies in the body that may lead to specific diseases. It also offers dosage recommendations tailored to different age groups, including infants, adults, and the elderly. Furthermore, the system evaluates the probability of side effects and presents detailed drug information, including physicochemical properties and 2D/3D structural representations. The model was trained and tested on relevant datasets, achieving an accuracy of 95%, demonstrating its effectiveness compared to conventional approaches. The results indicate that the system can significantly enhance the efficiency and reliability of drug discovery and recommendation processes. In conclusion, the proposed platform serves as a comprehensive and intelligent tool for supporting medical professionals, enabling faster diagnosis support, personalized treatment planning, and improved patient outcomes.

Keywords: *Artificial intelligence; Drug discovery; Machine learning; Predictive modeling; Random forest*

1. Introduction

Drug discovery is a fundamental process in pharmaceutical research aimed at identifying effective therapeutic compounds for treating various diseases. However, the conventional drug development pipeline is highly time-consuming, expensive, and requires extensive experimental validation. It involves screening a large number of chemical compounds followed by multiple stages of safety and efficacy testing. These limitations highlight the need for efficient and data-driven approaches to accelerate the early stages of drug discovery. In recent years, advancements in machine learning and artificial intelligence (AI) have significantly transformed biomedical research by enabling the analysis of large-scale chemical and biological datasets[1]. These techniques help identify complex relationships between molecular properties and drug effectiveness, allowing researchers to predict potential drug candidates prior to laboratory validation. As a result, computational approaches can reduce both time and cost associated with traditional methods. Several studies have explored the application of machine learning in drug discovery (Johnson and Lee, 2022; Gupta and Sharma, 2023). These works demonstrate that data mining, classification algorithms, and visualization techniques can improve prediction accuracy and support decision-making. However, most existing approaches focus on specific tasks such as prediction or classification, lacking a comprehensive framework that integrates multiple functionalities into a single system[2]. To address these limitations, this study proposes a machine learning-based drug discovery platform that combines data preprocessing, feature

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

extraction, predictive modeling, and visualization. Unlike existing methods, the proposed system provides an integrated and scalable solution for identifying potential drug candidates and generating meaningful insights. This approach aims to enhance the efficiency, accuracy, and reliability of drug discovery processes [3].

1. 1 Background Of The Study

The traditional drug discovery process relies heavily on laboratory experiments and manual analysis, making it both resource-intensive and time-consuming[5]. Researchers must evaluate numerous chemical compounds and their interactions with biological targets, which increases the complexity of the process. As highlighted in previous studies, handling large volumes of molecular data using conventional methods is inefficient and prone to errors. Machine learning provides an effective solution to these challenges by enabling automated data analysis and predictive modeling. It allows the system to learn from existing datasets and identify relationships between molecular features and drug effectiveness[6]. This shift from experimental to computational approaches has significantly improved the efficiency of pharmaceutical research Figure 1.

1.2 Objective Of The Study

The main objective of this study is to develop an intelligent drug discovery platform using machine learning techniques that can assist researchers in identifying potential drug candidates efficiently. The system focuses on analyzing chemical compound data, extracting relevant features, and applying machine learning models to predict biological activity[7].

The proposed work aims to:

- Reduce the time and cost involved in traditional drug discovery
- Improve prediction accuracy using machine learning models
- Provide a user-friendly platform for analyzing chemical datasets
- Support data-driven decision-making in pharmaceutical research

2. Method

The proposed methodology presents a structured machine learning framework designed to identify potential drug candidates from chemical compound datasets. The system follows a step-by-step approach that transforms raw molecular data into meaningful predictions [4]. Each stage in the pipeline is carefully designed to improve accuracy, reduce computational complexity, and ensure reliable outcomes.

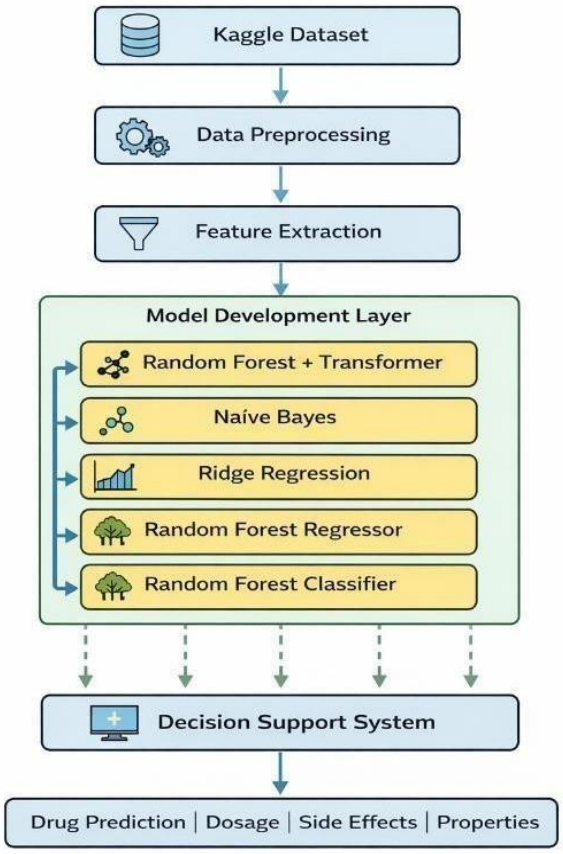

Figure 1 System Architecture of Proposed Model

2.1 Data Description

The proposed system utilizes multiple datasets, each containing approximately 14,000 records, to support different predictive tasks such as drug recommendation, side-effect prediction, deficiency identification, dosage estimation, and drug property analysis[8]. The drug recommendation dataset includes disease names, drug information, and features such as absorption, toxicity, and solubility, which are used for ranking and selection of suitable drugs. The side-effect dataset contains drug attributes and corresponding severity values, enabling

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

regression-based prediction evaluated using MAE and RMSE. The deficiency dataset consists of disease-related biological deficiencies and is processed using text encoding techniques for classification using Multinomial Naive Bayes. The dosage dataset includes drug names, age groups (infants, adults, elderly), and dosage values, which are used to train a Ridge Regression model. Additionally, the drug properties dataset contains physicochemical features such as lipophilicity and binding affinity, which are used for regression analysis. All datasets undergo preprocessing steps including data cleaning, encoding, and normalization to ensure consistency and improve model performance[9].

2.2 Data Preprocessing

Data preprocessing is a crucial step to ensure the quality and consistency of the datasets used in the proposed system. Initially, all datasets were examined for missing and inconsistent values. Missing data was handled using appropriate techniques such as removal or imputation to maintain dataset integrity. Categorical and textual data, particularly in the deficiency dataset, were converted into numerical format using encoding techniques suitable for Machine Learning models. Feature scaling and normalization were applied to numerical attributes such as absorption, toxicity, solubility, dosage, and physicochemical properties to improve model performance and convergence. Outliers were identified and treated to reduce their impact on prediction accuracy. The datasets were then divided into training and testing sets to evaluate model performance effectively. Additionally, feature selection techniques were applied to retain only the most relevant attributes, reducing dimensionality and improving computational efficiency[10]. These preprocessing steps ensure that the data is clean, structured, and suitable for training multiple Machine Learning models used in the system.

2.3 Feature Extraction

Feature extraction is performed to identify and select the most relevant attributes from the datasets for effective model training[11]. In the proposed system, different features are extracted based on the specific task. For drug recommendation, key pharmacokinetic features such as absorption, toxicity, and solubility are extracted and used for ranking drugs. In the side-effect prediction module, drug-related attributes are selected to estimate the probability and severity of side effects[12]. For deficiency identification, textual data related to diseases and deficiencies are transformed into numerical representations using feature extraction techniques such as vectorization, enabling classification using Multinomial Naive Bayes. In the dosage prediction module, features such as drug type and patient age group are used to estimate appropriate dosage values Figure 2. Additionally, physicochemical properties such as lipophilicity and binding affinity are extracted for drug property analysis. These features play a crucial role in understanding drug behavior and effectiveness. The extracted features are further refined through selection techniques to improve model accuracy and reduce computational complexity.

2.4 Model Training

2.4.1 Drug Recommendation

The drug recommendation module utilizes a Random Forest classifier combined with transformer-based techniques. Random Forest is an ensemble learning method that consists of multiple decision trees, where each tree is trained on a random subset of the data and features. The final prediction is obtained through majority voting across all trees.

Input features such as absorption, toxicity, and solubility are used to train the model. Transformer-based encoding is applied to capture complex relationships within the data. This hybrid approach improves the model's ability to rank and recommend suitable drugs for a given disease.

2.4.2 Side-Effect Prediction

The side-effect prediction module is implemented using a Random Forest Regressor, which extends the Random Forest structure for continuous output prediction. The model consists of multiple regression trees, where each tree predicts a numerical value, and the final output is obtained by averaging the predictions. Drug-related attributes are used as input features to estimate the likelihood and severity of side effects. The ensemble nature of the model helps in reducing overfitting and improving prediction

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

stability.

2.4.3 Deficiency Identification

The deficiency identification module uses Multinomial Naive Bayes, a probabilistic classifier based on Bayes' theorem. The model assumes feature independence and calculates the probability of each class given the input features. Textual data representing diseases and deficiencies are converted into numerical form using vectorization techniques. The model computes posterior probabilities for each class and selects the class with the highest probability, making it suitable for text-based classification tasks[13].

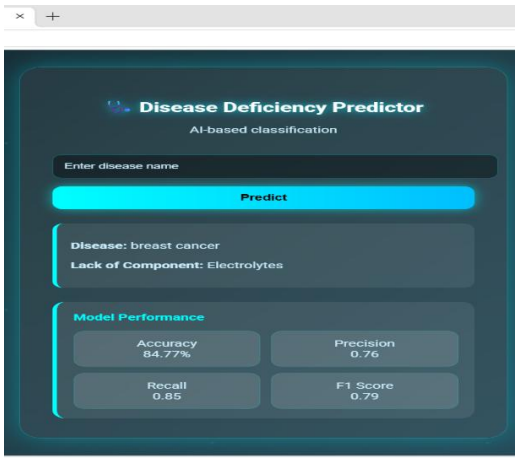

Figure 2 UI page of Deficiency Predictor

2.4.4 Dosage Prediction

The dosage prediction module employs Ridge Regression, a linear regression model with L2 regularization. The model minimizes the cost function by adding a penalty term to control large coefficient values and reduce overfitting.Input features include drug type and patient age categories[15]. The model learns the relationship between these inputs and the corresponding dosage values, producing stable and generalized predictions.

2.4.5 2D And 3D Representation of Drug

Visualization of drug structures is an important aspect of understanding their chemical and biological properties. In the proposed system, drug structures are represented in both two-dimensional (2D) and three-dimensional (3D) formats using the PubChemPy library[16].

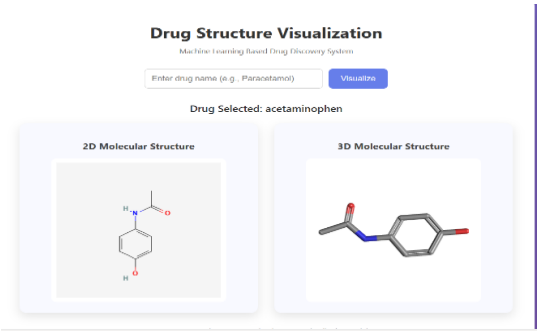

Figure 3 2D And 3D Drug Structure Visualization

2.4.6 Drug Property Prediction

The drug property prediction module uses a Random Forest Regressor to estimate physicochemical properties such as lipophilicity and binding affinity. Similar to other ensemble methods, multiple regression trees are trained, and their outputs are averaged to obtain the final prediction [14].

Table1 Machine Learning Models Used in System

Features	Machine Learning Model Used	Task
Drug Recommendation	Random Forest Classifier	Classification
Side effect Prediction	Random Forest Regressor	Regression
Deficiency Prediction	Naïve Bayes	Classification
Dosage Prediction	Ridge Regression	Regression
Drug Properties Prediction	Random Forest Regressor	Regression

2.4.7 Training Pipeline

Each module follows a structured training pipeline consisting of data preprocessing, feature extraction, model training, and validation. Separate training scripts are implemented for each dataset to ensure modularity and efficient execution. This design allows independent optimization of models and improves the overall performance and scalability of

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

the system Figure 3.

3. Results

The performance of the proposed AI-driven drug discovery platform was evaluated through a series of experiments conducted on multiple datasets, each containing approximately 14,000 records. The datasets were divided into training and testing sets to ensure unbiased evaluation. Different Machine Learning models were applied to specific tasks, and appropriate evaluation metrics were used based on the nature of each problem. For the drug recommendation module, a Random Forest classifier combined with transformer-based techniques was used to rank drugs based on features such as absorption, toxicity, and solubility. The model achieved an accuracy of 95%, indicating its effectiveness in identifying suitable drugs. The side-effect prediction module was implemented using a Random Forest Regressor. The experiment focused on predicting the severity of side effects based on drug-related attributes Table 1. The model achieved a Mean Absolute Error (MAE) of 1.49 and a Root Mean Square Error (RMSE) of 1.71, demonstrating reliable regression performance. For deficiency identification, the Multinomial Naive Bayes model was trained on encoded textual data. The experiment aimed to classify diseases based on associated biological deficiencies. The model achieved an accuracy of 84.77% and a recall of 0.85, indicating good classification capability Table 2. The dosage prediction module utilized Ridge Regression to estimate appropriate dosage values for different age groups[17]. The model achieved a Mean Absolute Error (MAE) of 0.699 and a Mean Squared Error (MSE) of 0.77, showing effective prediction of continuous values. The drug property prediction module employed a Random Forest Regressor to estimate properties such as lipophilicity and binding affinity. The model achieved a Mean Absolute Error (MAE) of 1.494 and a Root Mean Square Error (RMSE) of 1.705, indicating consistent performance. Overall, the experimental results demonstrate that the proposed system effectively handles multiple tasks and provides accurate and reliable outputs across different modules.

Table 2 Results of Classification Models

Machine Learning Model Used	Task	Accuracy	Precision	Recall	F1 Score
Random Forest Classifier	Drug Recomme-ndation	95	92	93	93
Multinomial Naive Bayes	Deficiency Prediction	84.77	76	85	79

Table 3 Results of Regression Models

Machine Learning Model Used	Task	Mean Absolute Error	Mean Square Error	Root Mean Square Error
Random Forest Regressor	Side Effects Prediction	1.49	2.92	1.71
Ridge Regression	Dosage Prediction	0.699	0.77	0.88
Random Forest Regressor	Drug properties Prediction	1.49	2.90	1.70

4. Discussion

The experimental results highlight the effectiveness of using a multi-model approach for drug discovery tasks. Instead of relying on a single model, the proposed system assigns specific Machine Learning techniques to different functionalities, which improves overall performance and flexibility. This modular design allows each model to focus on its strengths, leading to better generalization across tasks. The strong performance of the drug recommendation module indicates that ensemble-based approaches are well-suited for handling complex relationships between pharmacokinetic features[20]. The combination of Random Forest and transformer-based techniques enables the system to capture both structured and contextual patterns, resulting in more reliable recommendations. The side-effect and drug property prediction modules

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

demonstrate the capability of regression models to capture continuous relationships within the data. The use of ensemble regression further enhances stability and reduces the impact of noise, making the predictions more consistent and dependable[18]. The deficiency identification module shows that probabilistic models like Multinomial Naive Bayes are effective for text-based classification tasks. Its ability to handle high-dimensional categorical data makes it suitable for identifying disease-related deficiencies Table 3. The dosage prediction module highlights the importance of regularization techniques in handling real-world medical data. Ridge Regression helps control overfitting and ensures stable predictions across different patient categories. One of the key contributions of this work is the integration of multiple functionalities into a single platform. Unlike traditional systems that focus on isolated tasks, the proposed approach provides a comprehensive solution that supports various aspects of medical decision-making. This integration improves usability and reduces the need for multiple independent tools. However, the system's performance depends on the quality and diversity of the datasets. Limited or biased data may affect prediction reliability. Future work can focus on incorporating larger datasets, real-time clinical data, and advanced deep learning models to further enhance system performance. Overall, the proposed system demonstrates significant potential in improving the efficiency and practicality of AI-driven drug discovery[19].

Conclusion

This study addressed the challenges associated with traditional drug discovery methods, which are often time-consuming, costly, and inefficient. The proposed AI-driven drug discovery platform successfully integrates multiple Machine Learning models to provide a comprehensive solution for healthcare applications. Based on the results and discussion, the system effectively performs various tasks, including drug recommendation, side-effect prediction, deficiency identification, dosage estimation, and drug property analysis. The modular approach enables each model to contribute efficiently to its respective task, improving overall system performance. The findings confirm that the proposed system enhances the efficiency and reliability of drug discovery processes by leveraging data-driven techniques. Additionally, the integration of multiple functionalities into a single platform supports better decision-making for healthcare professionals. Thus, the proposed approach provides a scalable and effective solution for modern drug discovery and demonstrates the potential of Artificial Intelligence in transforming healthcare systems.

Acknowledgements

We sincerely thank Rajiv Gandhi University of Knowledge Technologies, R.K. Valley, Kadapa, and the Department of Computer Science and Engineering for their constant support and encouragement during this project. We are also grateful to our faculty members for their valuable guidance and motivation throughout the completion of this research work.

Authors Biographies

G. Swapna is a lecturer in the Department of Computer Science and Engineering at RGUKT, R.K. Valley, with more than nine years of teaching experience. Her academic interests span software engineering, big data, computer networks, and web technologies, along with related emerging areas. E. Susmitha is an Assistant Professor in the Department of Computer Science and Engineering at RGUKT, R.K. Valley, and is pursuing her Ph.D. at JNTUA College of Engineering, Anantapuramu. With over nine years of teaching experience, she has authored a book, obtained a patent, published around eleven international journal papers, presented work at international conferences, and participated in numerous faculty development programs. She received the Prathibha Award in M.Tech in 2016 and has qualified both UGC-NET and APSET. Her research interests include machine learning, big data, mobile applications, web technologies, the Internet of Things, and artificial intelligence

References

[1]. Vamathevan, J., Clark, D., Czodrowski, P., Dunham, I., Ferran, E., Lee, G., Li, B., Madabhushi, A., Shah, P., Spitzer, M., & Zhao, S. (2019). Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6), 463–477. https://doi.org/10.1038/s41573-019-0024-5

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

[2]. Zhavoronkov, A., Ivanenkov, Y. A., Aliper, A., Veselov, M. S., Aladinskiy, V. A., Aladinskaya, A. V., Terentiev, V. A., Polykovskiy, D. A., Kuznetsov, M. D., Asadulaev, A., Volkov, Y., Zholus, A., Shayakhmetov, R., Zhebrak, A., Minaeva, L. I., Zagribelnyy, B. A., Lee, L. H., Soll, R., Madge, D., … Aspuru-Guzik, A. (2019). Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nature Biotechnology*, 37(9), 1038–1040. https://doi.org/10.1038/s41587-019-0224-x

[3]. Paul, D., Sanap, G., Shenoy, S., Kalyane, D., Kalia, K., & Tekade, R. K. (2021). Artificial intelligence in drug discovery and development. *Drug Discovery Today*, 26(1), 80–93. https://doi.org/10.1016/j.drudis.2020.10.010

[4]. Chen, H., Engkvist, O., Wang, Y., Olivecrona, M., & Blaschke, T. (2018). The rise of deep learning in drug discovery. *Drug Discovery Today*, 23(6), 1241–1250. https://doi.org/10.1016/j.drudis.2018.01.039

[5]. Ekins, S. (2016). The next era: Deep learning in pharmaceutical research. *Pharmaceutical Research*, 33(11), 2594–2603. https://doi.org/10.1007/s11095-016-2029-2

[6]. Mayr, A., Klambauer, G., Unterthiner, T., & Hochreiter, S. (2018). DeepTox: Toxicity prediction using deep learning. *Frontiers in Environmental Science*, 3, 80. https://doi.org/10.3389/fenvs.2015.00080

[7]. Kearnes, S., McCloskey, K., Berndl, M., Pande, V., & Riley, P. (2016). Molecular graph convolutions: Moving beyond fingerprints. *Journal of Computer-Aided Molecular Design*, 30(8), 595–608. https://doi.org/10.1007/s10822-016-9938-8

[8]. Altae-Tran, H., Ramsundar, B., Pappu, A. S., & Pande, V. (2017). Low data drug discovery with deep learning. *ACS Central Science*, 3(4), 283–293. https://doi.org/10.1021/acscentsci.6b00367

[9]. Feinberg, E. N., Sur, D., Wu, Z., Husic, B. E., Mai, H., Li, Y., Sun, S., Yang, J., Ramsundar, B., & Pande, V. (2018). PotentialNet for molecular property prediction. *ACS Central Science*, 4(11), 1520–1530. https://doi.org/10.1021/acscentsci.8b00507

[10]. Jiménez-Luna, J., Grisoni, F., & Schneider, G. (2020). Drug discovery with explainable AI. *Nature Machine Intelligence*, 2(10), 573–584. https://doi.org/10.1038/s42256-020-00236-4

[11]. Yang, K., Swanson, K., Jin, W., Coley, C., Eiden, P., Gao, H., Guzman-Perez, A., Hopper, T., Kelley, B., Mathea, M., Palmer, A., Settels, V., Jaakkola, T., Jensen, K., & Barzilay, R. (2019). Analyzing learned molecular representations. Journal of Chemical Information and Modeling, 59(8), 3370–3388. https://doi.org/10.1021/acs.jcim.9b00237

[12]. Stokes, J. M., Yang, K., Swanson, K., Jin, W., Cubillos-Ruiz, A., Donghia, N., MacNair, C. R., French, S., Carfrae, L. A., Bloom-Ackermann, Z., Tran, V. M., Chiappino-Pepe, A., Badran, A. H., Andrews, I. W., Chory, E. J., Church, G. M., Brown, E. D., Jaakkola, T. S., Barzilay, R., & Collins, J. J. (2020). A deep learning approach to antibiotic discovery. Cell, 180(4), 688–702. https://doi.org/10.1016/j.cell.2020.01.021

[13]. Schneider, G. (2018). Automating drug discovery. Nature Reviews Drug Discovery, 17(2), 97–113. https://doi.org/10.1038/nrd.2017.232

[14]. Baskin, I. I. (2020). Machine learning methods in computational toxicology. Expert Opinion on Drug Metabolism & Toxicology, 16(9), 757–764. https://doi.org/10.1080/17425255.2020.1795261

[15]. Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning in bioinformatics. Briefings in Bioinformatics, 20(6), 2182–2200. https://doi.org/10.1093/bib/bby094

[16]. Gawehn, E., Hiss, J. A., & Schneider, G. (2016). Deep learning in drug discovery. Molecular Informatics, 35(1), 3–14. https://doi.org/10.1002/minf.201501008

International Research Journal on Advanced Engineering and Management
https://goldncloudpublications.com
https://doi.org/10.47392/IRJAEM.2026.0273
e ISSN: 2584-2854
Volume: 04 Issue: 05
May 2026
Page No: 1863 - 1870

[17]. Sliwoski, G., Kothiwale, S., Meiler, J., & Lowe, E. W. (2014). Computational methods in drug discovery. Pharmacological Reviews, 66(1), 334–395. https://doi.org/10.1124/pr.112.007336

[18]. Dahl, G. E., Jaitly, N., & Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. Journal of Chemical Information and Modeling, 54(4), 1123–1132. https://doi.org/10.1021/ci4007009

[19]. Korotcov, A., Tkachenko, V., Russo, D. P., & Ekins, S. (2017). Comparison of deep learning with multiple machine learning methods. Molecular Pharmaceutics, 14(12), 4462–4475. https://doi.org/10.1021/acs.molpharmaceut.7b00578

[20]. Walters, W. P., & Murcko, M. (2020). Assessing the impact of generative AI on medicinal chemistry. Nature Biotechnology, 38(2), 143–145. https://doi.org/10.1038/s41587-020-0418-2