International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Artificial Intelligence Assisted Drug Discovery of Noncommunicable Disease: Predictive Modelling and Optimization

Ayush Patel, Sangeeta Vhatkar, Namdeo Badhe
Dept. of Information Technology Thakur College of Engineering & Technology
(of Affiliation) University of Mumbai

Abstract- AI and machine learning are shaping up drug discovery and it is about time. The old way- slow, expensive and full of dead-ends- are outdated. Tools like deep learning, graph neural networks, GANs and reinforcement learning are stepping up. These tools actually help scientists spot new targets, sift through virtual libraries for promising compounds, predict how molecules will behave, dream up brand new drug designs, find fresh uses for old drugs and even streamline clinical trials. Graph models, in particular, shine because they get the complicated shape and connections in molecules. These all let researchers simulate how tiny structures interact in the messy reality of biology. Generative AI pushes boundaries even further by designing all sorts of molecules- each tailored for certain properties- across an almost endless chemical universe. Technology is making and creating waves everywhere: cancer, heart conditions, brain disorders, infections-you name it. Across the board, the results are better predictions, smarter trade-offs, more molecular variety and a smoother path from lab to clinic. Of course, it's not all smooth sailing. Challenges remain like messy data, black-box designing making, regulatory headaches and the tricky business of converting code into medicine. But even with those bumps, AI-powered drug discovery isn't another upgrade. It is a real-shift: more data-driven, more scalable and a lot more personal. The evidence keeps piling up-AI is speeding up therapeutic breakthroughs and rewriting the future position of medicine, one algorithm at a time.

Keywords- Artificial Intelligence, Drug Discovery, Non-communicable diseases, Ensemble Learning, Feature Engineering, Biomedical Data Analysis, Disease Prediction, Stacking Model.

I. INTRODUCTION

Background

Finding new drugs take a lot of time and money. From the first idea to getting regulatory approval, the process usually drags on for over ten years and most drug candidates never make it past the clinical trials [1].[2],[3]. Part of the problem is that biology isn't simple-disease work through tangled networks and can look very different from one patient to another. On top of that, there's this huge world of possible compounds out there, most of it still unexplored, which makes picking out safe and effective drug candidates feel like searching needle in a haystack [4],[5]. Sure, some traditional lab techniques and high-output screening help, but they can eat up a lot of resources and can only cover so much ground. All these hurdles keep pushing costs higher and slow everything down. That is why there is a stronger push for smarter, data-driven strategies to make drug discovery faster and cheaper. [6],[7]

Emergence of AI in drug discovery

AI and machine learning are changing the game when it comes to handling massive datasets of geonic, proteomic, chemical and clinical data. With powerful algorithms, they make disease modeling sharper, spot new drug targets faster and even speed up virtual screening [6],[7],[8],[9]. The real magic is how they shift through humongous data and find patterns that humans would probably miss. This means smarter decisions at every step of drug discovery. In the end, teams using AI and ML get drugs to market quicker, spend less money and also see better results. [9],[10]

Advanced AI methodologies

Lately, deep learning models have taken off in drug discovery. GANs, Autoencoders, GNNs and so on. These tools help researchers dream up brand new molecules from scratch, predict how tightly a drug will bind to its target, estimate

properties of different compounds and even find new uses for existing drugs [11],[12],[13],[14]. Graph based models stand out here. They map molecules in a way that feels almost obvious, using graphs to show how atoms connect and how they're arranged in space. Because of this, they do a better job at capturing the real structure of molecules and how they behave. This approach boosts accuracy and makes predictions more reliable, even when researchers throw in all sorts of different chemical and biological data. In the end, these advances are making AI a much stronger player in the world of pharmaceutical research [15],[16].

Scope and Significance

AI is making a real difference in how we tackle everything from heart disease and brain disorders to cancer and infectious diseases. Researchers are finding that these new tools speed up discovery, let them juggle multiple goals at once, and help move ideas from the lab to the clinic more smoothly. What's really impressive is how AI can pull together tons of different biomedical data, fine-tune drug properties all at once, and open the door to more personalized treatments [17-28]. Sure, there are still bumps in the road—like making sure we really understand how the models work, getting everyone on the same page with data, and keeping up with regulations. But even with these hurdles, AI is changing the game. Pharmaceutical research is becoming more about data, precision, and scaling up smart solutions that focus on accurate predictions and treatments tailored to each patient [17-28].

II. LITERATURE REVIEW

Table 2.1

References	Title	AI Methodology	Key Findings	Limitations
[1]	AI in Drug Discove ry and Delivery	ML/DL	Overvie w of AI across formulati on and optimization	Lacks empirical impleme ntation details
[2]	AI- Aided Drug Discove ry Pipeline	ML	Highlight s preclinic al risk reduction through AI	Broad framewo rk, less technical specificit y
[3]	GNNs in AI-aided Drug Discove ry	GNN	Deep dive into scalable and explaina ble GNNs	High-level overview , lacks real-world validation
[4]	Feedbac k GAN for Molecule Design	GAN + NSGA	Generate s novel, high-diversity molecule s with stereoche mical info	Needs broader experime nt validational
[5]	AI for Cardiov ascular Drug Develop ment	ML	Domain-specific applicati on in cardiovas cular therapeutics	High-level, lacks algorith mic focus
[6]	GNN Advanc es in Drug Discovery	GNN	Categori zes GNN types and benchma rk datasets	Overvie ws only post 2021 developments
[7]	DeepDr ug: Graph Learnin g Frame work	GNN + CNN	Predicts drug–target interactio ns effectivel y	Performa nce depends on data quality
[8]	Transfor mer Models for QSAR	Transformer Models for	SMILES -based molecule property	Limited interpreta bility of Transfor mer

		QSAR	predictio n	models
[9]	Reinforc ement Learnin g in Drug Discovery	RL	Explores RL for optimizin g molecula r features	Constrai ned by search space complexi ty
[10]	Multi- modal AI	CNN + Bioinformatic s	Integrati on of bio-data	Limited in chemical
	Optimiz ation		for better predictio ns	synthesis capabilit es

Table 2.2

Paper No.	Title	Algorithm Used	Key Gaps Identified
[11]	Transforming Cardiovascular Care with Artificial Intelligence: From Discovery to Practice	ML, DL, Multimodal AI	Limited real-world validation; bias & workflow integration issues
[12]	AI-driven Drug Discovery and Repurposing Using Multi-omics for Myocardial Infarction and Heart Failure	Explainable AI (XAI), Multi-omics models	Data heterogeneity; limited translational validation
[13]	Deep Learning-Based Modeling of Drug–Target Interaction Prediction Incorporating Binding Site Information	CNN (DeepPS), SMILES + Binding motifs	Limited 3D structural integration; generalization concerns
[14]	Artificial Intelligence-Assisted Drug and Biomarker Discovery for Glioblastoma: A Scoping Review of the Literature	ML, DL	Lack of large-scale validation; reproducibility issues
[15]	Graph Neural Networks in Modern AI-aided Drug Discovery	GNN, MPNN, GCN, GAT	Oversmoothing; scalability; interpretability challenges
[16]	Application of Artificial Intelligence (AI) in Pharmaceutical Industry: In-Depth Review	ML, DL, NLP	Ethical concerns; data privacy; regulatory barriers
[17]	Artificial Intelligence in Drug Discovery	ML, Virtual Screening	High attrition rate; limited predictive accuracy
[18]	Artificial Intelligence Revolution in Drug Discovery: A Paradigm Shift in Pharmaceutical Innovation	ML, DL, NLP, AlphaFold-based tools	Data integration & interpretability issues
[19]	A Review on Machine Learning Approaches and Trends in Drug Discovery	QSAR, SVM, RF, ANN, CNN, GNN	Descriptor dependency; lack of standardization

[20]	GraphDTA: Predicting Drug–Target Binding Affinity with Graph Neural Networks	GCN, GAT, GIN (GNN-based)	Protein representation limited to sequences; no full 3D modeling

TABLE 2.3

Paper No	Title	AI Approach	Target Focus	Primary Data Type
[21]	Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine	ML, DL, AI workflows	Drug discovery & delivery	Genomic, proteomic & clinical datasets
[22]	Relevant Applications of Generative Adversarial Networks in Drug Design and Discovery	GAN-based Deep Learning	Molecular de novo design	Chemical structures & scRNA-seq data
[23]	How Generative Artificial Intelligence Can Transform Drug Discovery?	Generative AI (GAN, VAE, Transformer)	Drug discovery R&D	Multi-omics & biological text data
[24]	Applications of Artificial Intelligence in Drug Repurposing	ML, DL, Virtual Screening	Drug repurposing	Drug-target interaction databases
[25]	PLASMOpred: A Machine Learning-Based Web Application for Predicting Antimalarial Small Molecules	ML (RF, GBM, SVM, CatBoost)	Antimalarial inhibitor prediction	PubChem bioassay & molecular fingerprints
[26]	Artificial Intelligence for Drug Discovery: Are We There Yet?	ML, Generative Chemistry	Small-molecule drug discovery	Pharmacodynamic & pharmacokinetic data
[27]	Recent Advances and Application of Generative Adversarial Networks in Drug Discovery, Development, and Targeting	GAN-based Models	Drug design & targeting	Chemical & biological datasets
[28]	AI-Driven Drug Discovery: Breaking Barriers in Pharmaceutical Research and Development	ML, DL, NLP	Pharmaceutical R&D optimization	Biological data & clinical records

III. METHODOLOGY

Data Collection and Integration

To conduct research, scientists have gathered science-based dataset collections of medical-related, non-communicable diseases such as breast and heart disease and diseases of the cardiovascular system, from publicly available machine-learning repositories and biomedical data repositories. Every science-based dataset holds multiple medical attributes representative of disease conditions and related biological indicators. For purposes of a unified analysis, the individual datasets were integrated into one combined dataset. Doing so allows predictive models to identify patterns based on multiple disease types, helps to enhance the generalization ability of the machine-learning system, is critical for creating effective AI-based models of healthcare, and is being extensively used in AI drug discovery research [1,2,17].

Data Preprocessing

The datasets were preprocessed after collecting them to improve their quality and reliability. Preprocessing included removing duplicate records, addressing missing values, and normalizing the input feature values. Thus, the above-defined actions assist in reducing inconsistencies in the input to ensure optimal model performance. The dataset was also assessed for class imbalance and the presence of superfluous attributes. In

machine-learning workflows, data preprocessing is a critical process since clean, structured input data will improve training and predictive accuracy. Prior works have shown that proper data preprocessing and modeling will appreciably increase the effectiveness of artificial intelligence (AI) models in biomedical research and pharmaceutical protocols [4], [20].

Feature Engineering

In order to improve the accuracy of predictions, new biomedical indicators were generated from the current set of variables using feature engineering methods. The generated features revealed hidden associations among biological characteristics and gave machine learning model additional data points. Ratio and composite measures are examples of generated features, which derive from medical measurements. Generated features are useful for illustrating the overall health of patients and enhancing predictive models' capacities to identify disease patterns. The use of feature engineering is generally accepted as an important technology to improve machine learning performance in healthcare and drug discovery applications [5], [17].

Machine Learning Model Implementation Various machine learning models have been employed to evaluate a biomedical dataset and determine disease results. Algorithms studied for this effort are:

- ExtraTrees Classifier
- LightGBM Classifier
- Cat Boost Classifier

All three of these algorithms are 'ensemble' machine learning algorithms that can learn complex nonlinear interactions in data sets associated with biomedicine. The purpose of ensemble models is to improve the accuracy of prediction and minimise the use of training data through the use of multiple decision trees. In recent studies, ensemble learning has been demonstrated to be an effective method for carrying out health care analytics and using AI during the drug discovery process because these algorithms are able to process large datasets efficiently [2],[20].

Model Evaluation

The evaluation of machine learning models was done using a number of classification metrics including:

- Accuracy
- Precision
- Recall
- F1-score

Also included were confusion matrices, which were analyzed to identify where the models made prediction errors in their classification performance. The use of these various methods of evaluation helps to determine the degree of reliability of predictive models and how consistently they perform across multiple datasets.The proper evaluation of models is important to the validation of an AI-enabled healthcare system and to provide confidence in the validity of predictions for use in biomedical research [17], [20].

IV. PROPOSED WORK

Proposed Framework

The proposed work proposes a predictive framework using artificial intelligence to analyze biomedical datasets and identify correlating patterns to non-communicable diseases. The proposed predictive framework will utilize machine learning and biomedical data analysis together to help predict diseases and aid with drug discoveries. Artificial intelligence will change the way that pharmaceutical research works at a high rate of speed through its ability to analyze vast amounts of biomedical data rapidly, as well as by increasing the ability to find therapeutic targets within the data. [1],[2].

Biomedical Data Integration and Feature Extraction

The suggested system combines several separate datasets in order to create a single record of all available data on a particular disease or condition. The new combined record gives researchers, clinicians and other health care providers access to a larger amount of statistical information about the biology of diseases than they would have otherwise had. By understanding more about the biological characteristics of a disease, they can create better models that use machine learning to identify patterns of behavior. Also, numerical representations of other biomedical variables are generated by using various methods of feature extraction. These features provide the necessary input for machine learning algorithms to perform analyses of correlations and relationships among different biological indicators. These types of data-driven methods have been extensively utilized in AI-enhanced biomedical research to enhance predictions of model performance.

Predictive Modeling using Ensemble Learning

The Predictive Modeling stage is designed to collect and analyze the data using several different types of Ensemble Machine Learning (ML) algorithms including ExtraTrees, LightGBM,and CatBoost in order to forecast the risk or presence of a specific disease condition. The Ensemble Model Algorithms were specifically chosen for their abilities to handle very complex datasets and provide very high levels of predictive accuracy; they accomplish this through the use of combined Decision Trees which in turn enhances robustness and reduces predictive errors. Ensemble ML applications have continued to show excellent performance rates on Biomedical Data Analysis, as well as predictive models for Chronic Disease Prediction [2][20].

Stacking Ensemble Architecture

An architecture called stacking ensemble is used in employ a number of base model predictions made by multiple different base models combined together through a metaclass algorithm to further enhance the prediction performance of the proposed design. The base models used in the stacking framework are ExtraTrees, LightGBM, and CatBoost, while Logistic Regression is the final estimator that represents the best predictions.

By using this architecture, the proposed system can exploit the individual strengths of various different types of machine learning algorithms and create a much higher level of predictive certainty. Ensemble methods based on stacking have been applied very successfully in healthcare analytics because they improve the stability of model performances as well as producing improved predictive performance [17], [20].

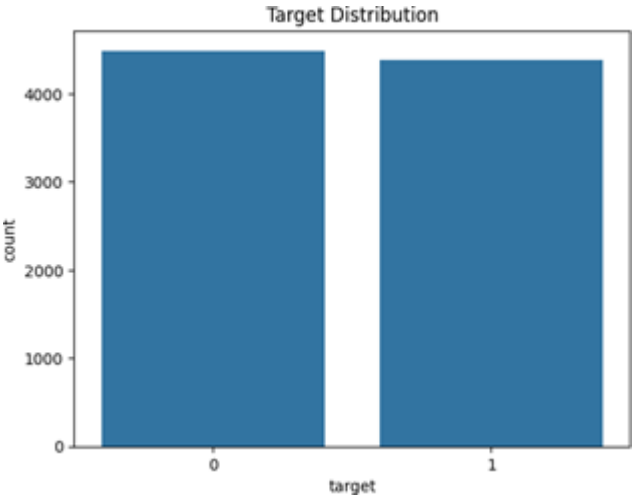
Disease Prediction and Drug Discovery Insights

The last step in the design of the suggested process will consist of data analysis (output) for making predictions and recognizable biomedical attributes (features or factors) that are critical to predicting disease occurrence. The analysis of feature importance determines what biological indicators have the most significant effect on the prediction result.

Identifying and characterizing these associations between feature(s), prediction results, and biological mechanism(s) is essential for providing insight into disease mechanisms and may aid researchers in discovering potential therapeutic targets for drug discovery. Thus, using AI to analyze biomedical data will help researchers identify potential targets for drug discovery early in the drug-discovery process, ultimately accelerating the overall drug-research process. AI is being used more frequently to discover innovative therapeutic prospects and enhance the efficacy of drug discovery methodologies.

V. RESULTS AND DISCUSSION

Dataset Distribution Analysis



Target Distribution Plot

The biomedical datasets used to evaluate the proposed AI drug design framework include many different types of biomedical data related to non-communicable diseases (e.g., breast cancer, heart disease, and all types of cardiovascular disease). These biomedical datasets were obtained from publicly available repositories and combined into one dataset to allow for predictions across multiple diseases.

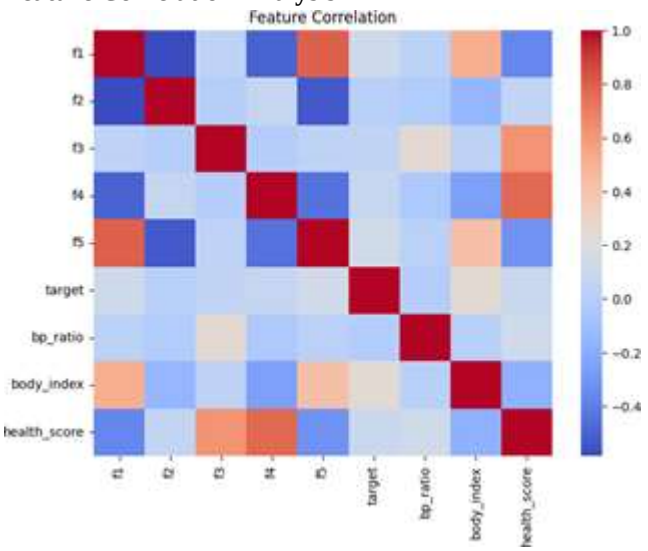
The combined dataset contains a diverse set of biomedical features (e.g., radius, texture, perimeter, and area measurements) that provide biological information that is important to the target or class labels (i.e., if the sample is from a diseased or healthy person).

In order to assess the degree of balance between the two classes of data (e.g., diseased and healthy), a target distribution analysis was conducted. The results of the target distribution analysis were displayed graphically showing the number of cases (positive and negative) for both diseased and healthy for the entire dataset. The proportion of classes (i.e., balance of the class) is one of the key considerations in building effective

predictive models for machine learning and in reducing bias in their predictions.

The analysis described above validates that the merged dataset is appropriate for developing disease detection and drug discovery predictive models [1],[2].

Feature Correlation Analysis

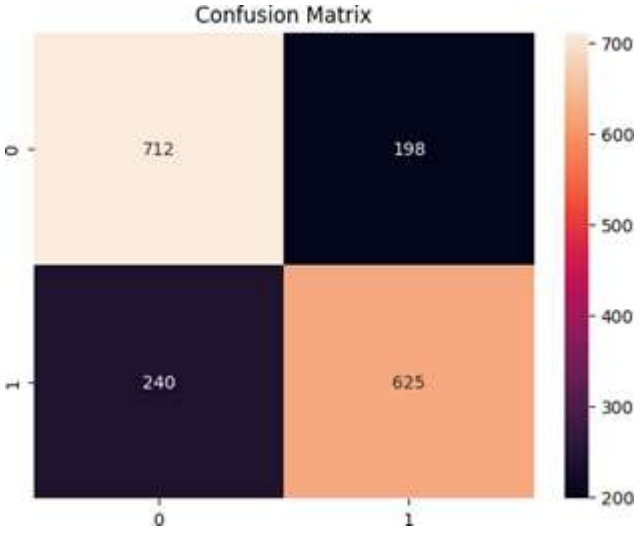


Feature Correlation Heatmap

Determining how biomedical variable relationships affect disease prediction will help correlate significant biological variables in advancing disease prediction through artificial intelligence. Analysis of correlations was performed using the Pearson Correlation Coefficient.

The correlation heatmap demonstrates how different biological variables relate to one another in terms of exerting influence on the prediction of disease. Variables that correlate positively or negatively indicate that a given variable is highly likely to affect a given disease outcome. For example, features that have a correlation between cell radius, cell perimeter, and cell area and disease outcomes have demonstrated strong correlations with their disease-related variables. Such relationships will help aid AI screening methods by determining the biological characteristics that are likely to be most predictive of predictive model performance. Correlation analysis will eliminate redundant variable selection and ultimately enhance model performance by helping to guide selection of the best variables to incorporate into the model [4], [17].

Confusion Matrix Analysis



Confusion Matrix

A Confusion matrix was produced to assess the classification performance of the stacking model. The Confusion matrix depicts all of the following categories of classification in terms of the actual versus predicted outcomes:

- True Positives – The number of correct disease predictions
- True Negatives – The number of correct healthy predictions
- False Positives
- False Negatives

The objective of a successful classification model is to maximize all true classifications while minimizing all false classifications.

The Confusion matrix indicates that the proposed model produces highly accurate results in identifying disease cases. Thus, the proposed model provides evidence of the ability to analyze predictive biomedical data accurately.

Evaluation techniques similar to the ones referenced in this document have been proven to be appropriate for evaluating classification performance in health care applications based on machine learning [2],[17].

Discussion of Results

Results of experimental studies suggest that biomedical data can be analyzed and the prediction of disease outcomes can occur with the use of machine learning methods. By combining

datasets of multiple diseases and engineered features, there is an increase in the robustness of the predictive model.

The stacking ensemble model was able to classify sample types effectively, proving the advantage of utilizing multiple algorithms for AI in the healthcare space. In addition, feature importance analysis helped to identify important biomedical variables that play a role in predicting a given disease. This supports the work of previous researchers who have identified an important role for AI in enhancing research in the biomedical field and improving existing systems that have been established to predict diseases. AI offers substantial potential for supporting the discovery of new drug therapies through the identification of biological patterns that are likely to have relevance when making decisions toward the identification of promising therapeutic targets [1], [2], [16].

VI. CONCLUSION

With 29 studies in front of us, one thing stands out: Artificial Intelligence is shaking up drug discovery in a big way. Researchers are using machine learning, deep learning, generative models, and graph models to zero in on drug targets, screen compounds faster, predict molecular properties, design new drugs from scratch, and even repurpose old ones. Graph Neural Networks and generative models deserve a special mention—they capture more detailed snapshots of molecules and help scientists sift through huge chemical spaces efficiently. Plus, they let researchers juggle things like effectiveness, safety, and how easy a drug is to make, all at once.

These studies dive into a range of diseases—heart problems, brain disorders, cancer, infections—and show just how much potential AI has to change the way we develop new treatments. Still, there are hurdles. Data quality is a big one, and so is figuring out why these models make the predictions they do. It's also tough to make sure an AI tool that works in a lab will work just as well in the real world, and then there's the question of meeting strict regulations. Even with these challenges, it's clear the field is moving away from old-school trial and error and toward smarter, data-driven drug development. For AI to truly deliver on the promise of faster, more personalized medicine, researchers need to push for better explainability, tougher benchmarks, and seamless end-to-end pipelines. That's how we unlock the real power of AI in medicine.

REFERENCES

1. A. U. Rehman et al., "Role of Artificial Intelligence in Revolutionizing Drug Discovery," Fundamental Research, 2025.
2. D. R. Serrano et al., "Artificial Intelligence (AI) Applications in Drug Discovery and Drug Delivery: Revolutionizing Personalized Medicine," Pharmaceutics, 2024.
3. S. Vatansever et al., "Artificial Intelligence and Machine Learning-Aided Drug Discovery in Central Nervous System Diseases: State-of-the-Arts and Future Directions," Medicinal Research Reviews, 2021.
4. X. Yang et al., "Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery," Chemical Reviews, 2019.
5. C. Cerchia and A. Lavecchia, "New Avenues in Artificial-Intelligence-Assisted Drug Discovery," Drug Discovery Today, 2023.
6. O. Zhang et al., "Graph Neural Networks in Modern AI-Aided Drug Discovery," 2025.
7. M. Abbasi et al., "Designing Optimized Drug Candidates with Generative Adversarial Network," Journal of Cheminformatics, 2022.
8. I. Kalansuriya et al., "AI-Driven Innovations in Cardiovascular Drug Development," Journal of Population Therapeutics & Clinical Pharmacology, 2024.
9. Z. Fang et al., "Recent Developments in GNNs for Drug Discovery," arXiv, 2025.
10. Khera et al., "Transforming Cardiovascular Care With Artificial Intelligence: From Discovery to Practice," 2024.
11. Sabry et al., "AI-Driven Drug Discovery and Repurposing Using Multi-Omics for Myocardial Infarction and Heart Failure," 2025.
12. D'Souza et al., "Deep Learning-Based Modeling of Drug–Target Interaction Prediction Incorporating Binding Site Information," 2023.
13. Conte et al., "Artificial Intelligence-Assisted Drug and Biomarker Discovery for Glioblastoma: A Scoping Review," 2025.
14. Ray Das et al., "Application of Artificial Intelligence (AI) in Pharmaceutical Industry: In-Depth Review," 2025.
15. Sudan et al., "Artificial Intelligence in Drug Discovery," 2020.

16. Jarallah et al., “Artificial Intelligence Revolution in Drug Discovery: A Paradigm Shift in Pharmaceutical Innovation,” 2025.
17. Carracedo-Reboredo et al., “A Review on Machine Learning Approaches and Trends in Drug Discovery,” 2021.
18. Nguyen et al., “GraphDTA: Predicting Drug–Target Binding Affinity with Graph Neural Networks,” 2020.
19. Tripathi et al., “Recent Advances and Application of Generative Adversarial Networks in Drug Discovery, Development, and Targeting,” Artificial Intelligence in the Life Sciences, 2022.
20. Hasselgren and Oprea, “Artificial Intelligence for Drug Discovery: Are We There Yet?” Annual Review of Pharmacology and Toxicology, 2024.
21. “Applications of Artificial Intelligence in Drug Repurposing,” 2025.
22. PLASMOpred: A Machine Learning-Based Web Application for Predicting Antimalarial Small Molecules,” 2025.
23. “Transforming Cardiovascular Care with Artificial Intelligence: From Discovery to Practice,” 2024.
24. “AI-Driven Drug Discovery and Repurposing Using Multi-Omics,” 2025.
25. “Deep Learning-Based Drug–Target Interaction Prediction,” 2023.
26. “Artificial Intelligence-Assisted Drug and Biomarker Discovery for Glioblastoma,” 2025.
27. “Graph Neural Networks in Modern AI-Aided Drug Discovery,” 2025.
28. “AI Driven Drug Discovery: Breaking Barriers in Pharmaceutical Research and Development,” 2020.