ISSN: 2773-7055
Proceedings of 3rd International Research Conference
Institute of Technology, University of Moratuwa
December 18, 2024

MACHINE LEARNING APPROACHES IN IN-SILICO DRUG DESIGN AND DEVELOPMENT: A COMPREHENSIVE REVIEW

S. M. Mahagama[1] and N. T. Jayatilake[2*]
[1]Institute of Technology University of Moratuwa, Sri Lanka
[2]Horizon Campus, Sri Lanka
sitharam@itum.mrt.ac.lk[1], nadunij@horizoncampus.edu.lk[2*]

ABSTRACT: Machine Learning (ML) is premised on the idea that machines can learn from data, recognize patterns, and make optimum decisions. Machine learning approaches include various algorithms for interpreting and gaining knowledge from data. Recently such ML-based techniques have been applied in drug development, bioinformatics, cheminformatics, and other areas of medicine as well. Drug design involves creating small molecules that are favorable in shape and charge to the biomolecular target to which they bind. Since experimental and lab procedures are limited in throughput, accuracy, and cost, they are unsuitable for broad deployment. Therefore, the development of in-silico target-drug designing methods has gained increasing attention globally due to the advantages of speed and low cost. In silico techniques in pharmaceutical designing are a type of computerized simulation that employs computer-aided technologies which initialize with an understanding of precise biochemical reactions within the body forming combinations of them to meet a therapy profile. Computerized methods provide the benefit of producing novel candidates for drugs faster and at lower prices. Virtual screening and de novo design, in silico ADME/T prediction, and improved methods for assessing protein-ligand interaction and structured-based drug design are the major functions of computational drug development. In-silico drug design refers to the use of computational methods and simulations to design and optimize drug candidates. This process involves steps such as Target Identification and Validation, Structure-Based Drug Design, Ligand-Based Drug Design, Virtual Screening, Molecular Dynamics Simulations, ADMET Prediction and Optimization. The adoption of ML algorithms in the search of medicines is applicable throughout this entire process. In this review article, the machine learning applications employed in In-silico drug design and discovery are explored in detail.

Keywords: computer-assisted drug design, drug discovery, in-silico drug design, machine learning

1. INTRODUCTION

Machine Learning (ML) approaches adopt various algorithms to teach machines to analyze data, recognize patterns, and make optimum decisions (Talevi et al., 2020). The limitations in throughput, lesser accuracy, and high cost in experimental procedures and lab equipment for traditional drug design have encouraged clinicians to increasingly adopt the Computer Assisted Drug Design (CADD) method, hence increasing global attention on in-silico target-drug designing methods, due to the advantages of speed and low cost. The in-silico drug design process involves steps such as Target Identification and Validation, Structure-Based Drug Design, Ligand-Based Drug Design, Virtual Screening, Molecular Dynamics Simulations, ADMET Prediction and Optimization (Rao, 2011).

This study addresses critical limitations in traditional drug design methods which are widely documented as being resource-intensive, costly, environmentally harmful, and ethically challenging due to high rates of chemical waste and reliance on animal testing (Rao, 2011; Dahl et al., 2014). With ML-based in-silico methods, recent research points to significant potential for faster, more sustainable, and economically feasible drug development processes, especially for underfunded areas like rare and neglected diseases (Rifaioglu et al., 2019; Patel et al., 2020). Despite these advancements, there is limited comparative analysis on efficiency, environmental benefits, and overall efficacy of specific ML techniques across drug design stages, a gap that this review aims to address. Therefore, the objective of the study is to analyze the benefits of machine learning methods in improving the drug design process focusing on how these methods enhance efficiency, sustainability, and precision of the drug discovery process.

2. METHODOLOGY

The methodology emphasizes the selection, classification, and assessment of studies from credible academic sources that demonstrate application of ML methods across seven critical phases of drug discovery to make the process efficient and effective (Shaker et al., 2021).The analysis aims to evaluate comparative effectiveness, benefits, and drawbacks of various ML approaches, including Deep Learning (DL), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machines (SVM), as they pertain to various phases of in-silico drug development. The initial literature search was conducted using several academic databases including PubMed, Google Scholar, and Scopus which are well-regarded for their comprehensive coverage of biomedical research.

Key terms such as "machine learning," "in-silico drug design," and "computational drug development" were used as a guide to ensure relevance of the research material. The focus was put on peer-reviewed articles conducted during past 10 years to ensure the timeliness of the findings. Research showing empirical findings, which includes assessments of predictive accuracy, computational effectiveness, and practical uses of machine learning methods throughout the stages of drug development (such as target identification, lead optimization, and ADMET prediction, among others) and research articles that examine the use of machine learning in key stages of drug development, such as virtual screening molecular docking, and structure-based drug design were the inclusion criteria for the study. Articles that do not contain experimental or empirical evidence, including those that are entirely theoretical without real-world confirmation, research that only examines traditional (non-computational) approaches to drug discovery or machine learning applications in unrelated domains, and investigations that present incomplete or unclear explanations regarding the machine learning utilized were excluded.

The initial search provided approximately 35 papers. Then the titles and abstracts of identified articles were reviewed to assess their significance. Full texts of potentially relevant studies were then examined to determine their suitability and adoptable content. Nearly 23 works were sorted out in full text for reference. The reference management software ‘Zotero’ was used to store organize and keep track of the referenced articles, and to make the documentation efficient. The evaluation of Machine Learning Algorithms process included comparing the effectiveness of machine learning methods across different parameters and phases of drug discovery based on key performance metrics such as prediction accuracy, precision, recall, computational efficiency, advantages and limitations.

3. RESULTS AND DISCUSSION

Table 1 summarizes how these methods address inefficiencies, environmental problems, and economic issues integral to traditional drug discovery processes. Deep Learning (DL) techniques, particularly Deep Neural Networks (DNNs) have demonstrated success in activity prediction and drug repurposing. To illustrate this, Dahl et al. utilized DNN with 2D topological descriptors on the Merck Kaggle testing data surpassing the Random Forest (RF) method (Dahl et al., 2022). Moreover, Mayr et al. in their Tox21 project applied multitask DNN models to predict toxic effects using HTS data, and forecasting drug indications in the work of Subramanian et al. and Aliper et al. seem to imply DNN's useful applicability and accuracy in drug repurposing initiatives (Aliper et al., 2016). RF which is termed a supervised algorithm is also applicable in high volume datasets, for example in RF's applications; Cano et al. rank and reduce features to classify data with high rates whilst reducing computation (Cano et al., 2017). Regarding breast cancer, Naïve Bayes (NB) algorithm was effectively used by Pang et al. in discriminating active from inactive estrogen receptor antagonists and displayed potential to be desirable drug leads (Pang et al., 2018). Lastly, Support Vector Machine (SVM) algorithm proves especially useful for virtual screening and compound scoring tasks; thus, Patel et al. employed SVM to assess the molecular interactions which contributed

greatly towards compound selection and the search of therapeutic potential (Patel et al., 2020). These case studies demonstrate the advantages of each of the machine learning methods in furthering drug design processes.

Table 1. Applications, advantages and limitations of ML techniques in in-silico drug designing.

ML Model	Application	Prediction Accuracy	Computational Accuracy	Advantages	Limitations	References
Deep Learning (DL)	Activity Prediction, QSAR Modeling	~90% (QSAR model)	High resource consumption; efficient with large datasets	High accuracy for large datasets, flexible for various applications	Complex model validation; issues with generalization to new data (overfitting with small datasets)	Dahl et al., 2014; Rifaioglu et al., 2019
Random Forest (RF)	Lead Optimization, Predicting Drug-Target Interactions	~85%	Moderate (scalable for large datasets but longer training times)	Handles noisy data well, works with multiple data types, high prediction accuracy	Higher memory and time costs during training	Breiman, 2001; Cano et al., 2017
Naïve Bayes (NB)	Predicting Ligand-Target Interactions, Early Stage Drug Discovery	~70-80%	High computational efficiency (performs well with small datasets)	Performs well with smaller, noisy datasets, easy to implement, quick training times	Assumes feature independence; limited ability to capture complex patterns or correlations in biological data	Pang et al., 2018; Patel et al., 2020
Support Vector Machine (SVM)	QSAR Modeling, Virtual Screening, Predicting Molecular Properties	~80-85%	Moderate to high (can handle high-dimensional data, but memory-intensive with large datasets)	Effective in high-dimensional spaces, provides reliable predictions for drug-target interactions with limited data	Slow training times with large datasets, high memory consumption during screening	Patel et al., 2020; Dara et al., 2022

4. CONCLUSION

ML approaches allow for effective, precise and environmentally friendly drug discovery processes. This paper demonstrates that the techniques, including Deep Learning (DL), Random Forest (RF), Naïve Bayes (NB), and Support Vector Machines (SVM) have their comparative advantage in certain aspects of the drug design process. For example, it is well known that DL methods, while being resource costly systems prone to overfitting small datasets, can reach close to 90% accuracy in QSAR model predictions. RF models demonstrate strong results in lead optimization and manage to withstand the onslaught of noisy data, achieving 85% predictive accuracy but require enough memory and time to train. NB algorithms, due to their much lower computational requirements, allow for early-stage interaction between ligands and targets to be determined, although their ability to engage in complex biological systems is rather limited. Finally, SVMs perform well in virtual

screening and compound scoring, managing high-dimensional data effectively though they also consume large amounts of resources when the dataset is expansive.

Researchers have used ML algorithms and techniques to derive potential drug candidates and in improving the properties of drug molecules and compounds. Hence, clinical testing has become more efficient, time and money saving leading to sustainable engineering aspects while identifying new medicines. In consequence, the whole process has become faster, efficient and more accurate. Incorporating such advanced engineering solutions into the in-silico drug design and development process causes reduced wastage that takes place during various medical experiments which will decrease the carbon footprint aligned with the drug design process when compared with the conventional approaches. With this review it can be concluded that integrating Machine Learning into drug development not only boosts advancements and innovation but also drives the process more towards greener practices making things eco-friendlier to achieve global sustainability.

5. REFERENCES

Aliper, A., Plis, S., Artemov, A., Ulloa, A., Mamoshina, P., & Zhavoronkov, A. (2016). Deep learning applications for predicting pharmacological properties of drugs and drug repurposing using transcriptomic data. *Molecular Pharmaceutics, 13*(7), 2524–2530. https://doi.org/10.1021/acs.molpharmaceut.6b00248

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cano, G., Garcia-Rodriguez, J., Garcia-Garcia, A., Perez-Sanchez, H., Benediktsson, J. A., Thapa, A., & Barr, A. J. (2016). Automatic selection of molecular descriptors using random forest: Application to drug discovery. *Expert Systems with Applications, 72*, 151–159. https://doi.org/10.1016/j.eswa.2016.12.008

Dahl, G. E., Jaitly, N., & Salakhutdinov, R. (2014). Multi-task neural networks for QSAR predictions. arXiv. https://doi.org/10.48550/arXiv.1406.1231

Pang, X., Weiqi, F., Wang, J., Kang, D., Xu, L., Zhao, Y., Liu, A., & Du, G.-H. (2018). Identification of estrogen receptor α antagonists from natural products via in vitro and in silico approaches. *Oxidative Medicine and Cellular Longevity, 2018*, 1–11. https://doi.org/ 10.1155/2018/6040149

Patel, L., Shukla, T., Huang, X., Ussery, D. W., & Wang, S. (2020). Machine learning methods in drug discovery. *Molecules, 25*(22), Article 22. https://doi.org/10.3390/molecules25225277

Rao, V., & Srinivas, K. (2011). Modern drug discovery process: An in silico approach.

Rifaioglu, A. S., Atas, H., Martin, M. J., Cetin-Atalay, R., Atalay, V., & Doğan, T. (2019). Recent applications of deep learning and machine intelligence on in silico drug discovery: Methods, tools and databases. *Briefings in Bioinformatics, 20*(5), 1878–1912. https://doi.org/10.1093/bib/bby061

Shaker, B., Ahmad, S., Lee, J., Jung, C., & Na, D. (2021). In silico methods and tools for drug discovery. *Computers in Biology and Medicine, 137*, Article 104851. https://doi.org/10.1016/j.compbiomed.2021.104851

Talevi, A., Morales, J. F., Hather, G., Podichetty, J. T., Kim, S., Bloomingdale, P. C., Kim, S., Burton, J., Brown, J. D., Winterstein, A. G., Schmidt, S., White, J. K., & Conrado, D. J. (2020). Machine Learning in Drug Discovery and Development Part 1: A Primer. *CPT: Pharmacometrics & Systems Pharmacology, 9(*3), 129–142. https://doi.org/10.1002/ psp4.12491