OPEN ACCESS

Pharmaceutical Communications

DOI: 10.55627/pharma.002.01.0297

Review Article

Advancing Pharmaceuticals with Machine Learning: A Short Review of Research and Development Applications

Wajeeha Khalid[1*], Muhammad Yahya Khalid[2], Madiha Hena[3], Atif Sarwar[1] and Shahzaib Iqbal[4]

[1]Department of Pharmaceutics, Shifa Tameer-e-Millat University, Islamabad, Pakistan
[2]Faculty of Engineering and Informatics, University of Bradford, Bradford, United Kingdom
[3]Department of Computing, Abasyn University Islamabad Campus, Islamabad, Pakistan
[4]Department of Electrical Engineering, Abasyn University, Islamabad, Pakistan
***Correspondence**: wajeeha.scps@stmu.edu.pk

Abstract

In recent years, the field of pharmaceutical research and development (RD) has seen a surge of interest in artificial intelligence (AI) and machine learning (ML) technologies. These advancements may transform the industry by addressing challenges related to data analysis, computational capabilities, and rising costs associated with drug development. ML techniques have been progressively refined and applied to various stages of drug discovery over the past 15–20 years. Notably, there is a growing focus on utilizing ML in clinical trial design, conduct, and analysis, which the COVID-19 pandemic and the increased reliance on digital technology in clinical trials have further accentuated. However, it is crucial to move beyond mere buzzwords and acknowledge that the scientific method remains essential for drawing meaningful insights from data. By doing so, we can distinguish between genuine advancements and exaggerated claims, leading to informed decision-making regarding the optimal integration of ML methods in drug development. This review aims to provide a comprehensive understanding of key concepts, understand real-world usage, and offer a balanced perspective on the effective utilization of ML in RD.

Keywords: Machine learning, drug development, precision medicine, probability of success, clinical trial design.

1. Introduction

Significant progress has been made in the fields of artificial intelligence (AI) and machine learning (ML) in recent years. These advancements have been fuelled by remarkable breakthroughs in computational technology, enabling more efficient collection and analysis of vast amounts of data. Concurrently, the pharmaceutical industry has faced escalating costs in bringing new drugs to market and delivering them to patients. In this review, the term' RD' will be used as a general reference encompassing the research, scientific exploration, and procedural aspects associated with the development of drugs (Micallef and Blin 2020), starting from initial drug discovery, clinical development(Keppel Hesselink 2018), conduct, and ultimately, life-cycle management. The process of developing a new drug is not only lengthy but also comes with a substantial financial burden, as evident from the following statistics: an average investment of $1.3 billion is required per drug (DiMasi, Grabowski, and Hansen 2016), whereas usually drug development time ranges from 5.9 to 7.2 years for non-oncology drugs. As for oncology, drugs take 13.1 years, with only a 13.8% success rate for all drug development programs leading to approval (Wong, Siah, and Lo 2018). Considering these challenges, the pharmaceutical industry has shown a growing

interest in the potential of ML techniques, primarily due to their automated nature, predictive capabilities, and a potential boost in efficiency. There is a clear need for both patient- and business-centered perspectives to streamline drug development, reduce costs, shorten development timelines, and increase the probability of success.

Over the last 15-20 years, ML approaches have been extensively used in drug development, demonstrating increasingly sophisticated advancements(Sharma et al. 2022). The application of machine learning, clinical trial design, operations, and analysis is developing as a contemporary area of positive disruption. Furthermore, the COVID-19 pandemic has accelerated the deployment of ML in clinical trials due to the greater reliance on digital technology for patient data collection. This study aims to provide a detailed examination of ML utilization in drug development and identify areas where it may have a significant impact. The goal is to provide a neutral perspective by distinguishing between too optimistic and overly gloomy expectations and promote the best implementation of machine learning in the field. The first part of the discussion will be an introduction to the essential concepts and terminology used in ML. Following this, an investigation into the most effective implementations of ML techniques in pharmaceutical research and development is carried out, with a particular emphasis on processing clinical trial data and a comparison with more traditional statistical methods based on inference. In addition to the applications highlighting the ongoing work in clinical trial methods, this article offers a comprehensive overview of the current state of machine learning in research and development. At the end of the article, the challenges and opportunities that lie ahead to fully realize ML's potential for research and development in the pharmaceutical industry are discussed.

1. Key Concepts and Terminologies of ML

This section takes an in-depth look at the major concepts and terminologies of ML. AI refers to a method for programming computers to think and act in humanlike ways. To develop AI skills, a subset of AI known as ML employs algorithms that are trained on input data. Deep learning (DL) is a branch of machine learning that use artificial neural networks to simulate the intricate structure of the human brain.

1.1. Machine Learning Techniques

ML approaches comprise a wide range of methods and techniques that allow computers to learn and make predictions/judgments without the need for explicit programming. These methods are collectively referred to as ML. These methods enable machines to analyze given data, detect different patterns, and extract feature insights, thereby improving their prediction skills. These strategies analyze complicated data patterns using statistical models and algorithms to derive significant insights for informed forecasts and decision-making.

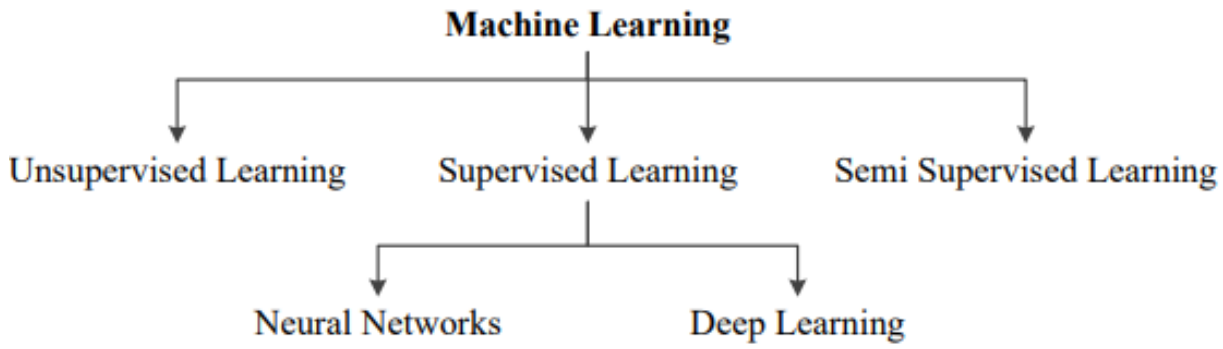

Figure 1: Machine Learning Techniques.

1.2. Unsupervised learning

The subfield of ML that focuses on using unlabeled data to uncover patterns that aid in solving clustering or association problems is known as 'unsupervised learning' (Iqbal, Khan, et al. 2022). Unlike supervised learning, unsupervised learning does not rely on labeled data with predefined outputs (Iqbal et al. 2023). Instead, it aims to find inherent structures and patterns within the data itself. Unsupervised learning algorithms analyze the input data to identify similarities, differences, or relationships among data points and then group or cluster them accordingly (Liu et al. 2023). The process of clustering aids in organizing and comprehending data, uncovering latent patterns or connections that may not be immediately evident. Unsupervised learning proves valuable in exploratory data analysis, anomaly detection, and data preprocessing tasks, providing a way to gain insights and perform analysis.

1.3. Supervised learning

The branch of ML that revolves around utilizing labeled datasets to train algorithms for precise data classification or outcome prediction is called 'supervised learning' (Iqbal, Naqvi, et al. 2022). In supervised learning, the algorithm learns from a dataset that already contains labeled instances, implying that the input-output pairs are known. By examining these labeled examples, the algorithm can discern patterns and correlations within the data, enabling it to accurately predict outcomes or classify unseen data instances. Supervised learning aims to train the algorithm to generalize from the labeled dataset and make accurate predictions on new unlabeled data.

1.4. Deep Learning (DL)

The field of supervised ML that employs multi-layered (deep) architectures to establish connections between inputs or observed features and outcomes is commonly referred to as 'deep learning.' Deep learning techniques use neural networks with several interconnected layers to learn complex patterns and representations from large datasets. Deep learning models can effectively collect and interpret complex relationships within data (Iqbal, Naqvi, et al. 2022) by leveraging these deep structures, permitting accurate predictions or classifications.

1.5. Neural Network (NN)

A highly parameterized model is a computer model with many configurable parameters. These models are frequently inspired by the complex network of neurons seen in the human brain, which is referred to as 'biological neural networks' (Yamazaki et al. 2022). These models seek to represent the complexity and nuances observed in biological systems by including a vast number of factors, allowing them to imitate

and recreate different elements of human brain activity(Nematzadeh et al. 2022).

1.6. Semi-Supervised Learning

A subset of machine ML called 'semi-supervised learning' involves combining a sizable amount of unlabeled data with little labeled data during the training phase. With the help of this approach, the algorithm can gain from both forms of data, improving performance(Yuhao Chen 2023). Unlabeled data offers insights into the underlying structure and trends of the data, whereas tagged data offers clear details regarding the anticipated outcomes. Starting with labeled data, the algorithm learns from the provided examples and their corresponding results. Even more patterns, resemblances, or correlations in the dataset are discovered using unlabeled data. The inclusion of untagged data enhances the algorithm's capacity to generalize and produce precise predictions. Semi-supervised learning becomes a suitable alternative when getting tagged data is expensive, difficult, or time-consuming. It is useful in fields where labeled data is in little supply, such as medical diagnosis, natural language processing, and picture recognition tasks, because it enables the use of a wealth of untagged data to enhance learning.

2. Using Machine Learning in Pharmaceutics

In this part, we present several cases illustrating the usage of ML techniques in pharmaceutical RD, highlighting its potential to enhance current procedures. Machine learning can assist scientists and researchers in increasing their RD projects' effectiveness, efficiency, and efficiency. Consequently, it may improve pharmacovigilance, clinical trial design, drug discovery, and therapy optimization.

3. Drug Discovery

ML techniques have been used to expedite drug discovery by accurately predicting novel pharmaceutical candidates' efficacy and safety (Li et al. 2015). ML models can suggest substances for further exploration by analyzing large amounts of molecular data, considerably decreasing the time and cost of traditional experimental procedures. This use of ML in drug discovery has the potential to speed up the identification and development of novel treatments, helping the pharmaceutical industry and improving patient outcomes.

4. Clinical Trial Design and Optimization

ML algorithms are improving clinical trial design and execution (Beck et al. 2020). By analyzing various patient data such as genetic profiles, demographics, and medical histories, these algorithms improve patient selection, identify acceptable objectives, and enable the deployment of adaptive trial designs. The application of ML methods in clinical trials improves efficiency and leads to more meaningful study results (Senior et al. 2019). This trial design and execution innovation can potentially speed up medical research while improving patient care.

5. Predictive Analytics for Patient Outcomes

Using ML models to predict treatment outcomes, disease progression, and potential adverse events, patient data may be effectively analyzed (John Jumper 2020) . These models integrate various data sources, such as electronic health records, genomic data, and real-world evidence (Beck et al. 2020). Through this integrated information, ML models assist clinicians in making personalized treatment decisions and optimizing patient care, ultimately resulting in

improved outcomes and tailored healthcare interventions.

6. Drug Repurposing and Combination Therapy

ML techniques offer immense potential in identifying opportunities for drug repurposing by exploring new therapeutic indications for existing drugs (Li et al. 2015). By analyzing extensive biomedical data encompassing drug-target interactions, molecular pathways, and disease associations, these techniques can uncover novel applications for approved drugs. Furthermore, they can identify synergistic drug combinations that could enhance treatment strategies. This approach not only allows for more efficient and cost-effective drug development but also employs existing knowledge and resources to drive innovation in the field of medicine.

7. Pharmacovigilance and Adverse Event Detection

ML models are actively being used to enhance pharmacovigilance efforts by efficiently identifying potential adverse events and drug interactions (Marcon, Queiroz, and Baquero 2022). These models analyze real-world data from various sources, including electronic health records, social media, and spontaneous reporting systems, to detect and monitor drug safety signals in a proactive and efficient manner. These application cases highlight the transformative power of ML technologies at various stages of pharmaceutical RD. ML techniques promote medication discovery and overall healthcare by boosting efficiency and precision and adopting patient-centric approaches.

8. Discussion and Conclusions

ML offers enormous growth potential in the realm of pharmaceutical RD. It does, however, bring challenges as well as opportunities. The use of digital technology in data collection has resulted in a significant increase in the volume and complexity of data. This expansion opens up new opportunities to improve our understanding of biological systems, repurpose medications, and provide valuable insights for clinical trial design and analysis in drug development. While recent developments in machine learning technologies are impressive, caution must be practiced to prevent the drawing of incorrect conclusions. It is crucial to consider confounding factors, employ reliable algorithms, analyze appropriate data, and fully comprehend the clinical questions underlying the endpoints and data collection. Proper training of ML algorithms is essential to ensure reliable performance in real-world scenarios with different data scenarios. Furthermore, not all research questions can be effectively addressed using ML, especially when faced with high variability, limited or poor-quality data, under-represented patient populations, or flawed trial design. The issue of under-represented patient populations is concerning, as it may introduce systematic bias. Additionally, as seen in other domains using ML, it is essential to exercise extreme care in addressing patient privacy and other bioethical considerations.

Conflict of Interest

The authors declare no conflict of interest.

Funding

NA.

Study Approval

NA.

Consent Forms

NA.

Authors Contribution

WK and SI conceptualized the study and wrote the final manuscript, MYK & WK helped in the analysis and writing the first draft, WK, AS, and MYK did the literature search and analysis, and SI supervised the whole project.

Acknowledgments

All the authors are thankful to the Shifa Tameer-e-Millat University for facilitating the writing of this review manuscript.

References

Beck, Bo Ram, Bonggun Shin, Yoonjung Choi, Sungsoo Park, and Keunsoo Kang. 2020. "Predicting commercially available antiviral drugs that may act on the novel coronavirus (SARS-CoV-2) through a drug-target interaction deep learning model." *Computational and Structural Biotechnology Journal* no. 18:784-790. doi: https://doi.org/10.1016/j.csbj.2020.03.025.

DiMasi, Joseph A., Henry G. Grabowski, and Ronald W. Hansen. 2016. "Innovation in the pharmaceutical industry: New estimates of R&D costs." *Journal of Health Economics* no. 47:20-33. doi: https://doi.org/10.1016/j.jhealeco.2016.01.012.

Iqbal, Shahzaib, Tariq M. Khan, Khuram Naveed, Syed S. Naqvi, and Syed Junaid Nawaz. 2022. "Recent trends and advances in fundus image analysis: A review." *Computers in Biology and Medicine* no. 151:106277. doi: https://doi.org/10.1016/j.compbiomed.2022.106277.

Iqbal, Shahzaib, Syed S. Naqvi, Haroon A. Khan, Ahsan Saadat, and Tariq M. Khan. 2022. "G-Net Light: A Lightweight Modified Google Net for Retinal Vessel Segmentation." *Photonics* no. 9 (12):923.

Iqbal, Shahzaib, Khuram Naveed, Syed S. Naqvi, Asim Naveed, and Tariq M. Khan. 2023. "Robust retinal blood vessel segmentation using a patch-based statistical adaptive multi-scale line detector." *Digital Signal Processing* no. 139:104075. doi: https://doi.org/10.1016/j.dsp.2023.104075.

John Jumper, Kathryn Tunyasuvunakool, Pushmeet Kohli, Demis Hassabis, and the AlphaFold Team. 2020. "Computational predictions of protein structures associated with covid-19." *Google DeepMind*.

Keppel Hesselink, Jan M. 2018. "Phenytoin repositioned in wound healing: clinical experience spanning 60 years." *Drug Discovery Today* no. 23 (2):402-408. doi: https://doi.org/10.1016/j.drudis.2017.09.020.

Li, Bin, Hyunjin Shin, Georgy Gulbekyan, Olga Pustovalova, Yuri Nikolsky, Andrew Hope, Marina Bessarabova, Matthew Schu, Elona Kolpakova-Hart, David Merberg, Andrew Dorner, and William L. Trepicchio. 2015. "Development of a Drug-Response Modeling Framework to Identify Cell Line Derived Translational Biomarkers That Can Predict Treatment Outcome to Erlotinib or Sorafenib." *PLOS ONE*

no. 10 (6):e0130700. doi: 10.1371/journal.pone.0130700.

Liu, Yajun, Dakui Wu, Wenju Zhou, Kefeng Fan, and Zhiheng Zhou. 2023. "EACP: An effective automatic channel pruning for neural networks." *Neurocomputing* no. 526:131-142. doi: https://doi.org/10.1016/j.neucom.2023.01.014.

Marcon, Danillo Silva, Mariana Ramos Queiroz, and Oswaldo Santos Baquero. 2022. "Adverse event classification and signal detection of data from the customer service and pharmacovigilance of a multinational veterinary pharmaceutical company." *Preventive Veterinary Medicine* no. 206:105704. doi: https://doi.org/10.1016/j.prevetmed.2022.105704.

Micallef, Joëlle, and Olivier Blin. 2020. "Orphan drug designation in Europe: A booster for the research and development of drugs in rare diseases." *Therapies* no. 75 (2):133-139. doi: https://doi.org/10.1016/j.therap.2020.02.003.

Nematzadeh, Sajjad, Farzad Kiani, Mahsa Torkamanian-Afshar, and Nizamettin Aydin. 2022. "Tuning hyperparameters of machine learning algorithms and deep neural networks using metaheuristics: A bioinformatics study on biomedical and biological cases." *Computational Biology and Chemistry* no. 97:107619. doi: https://doi.org/10.1016/j.compbiolchem.2021.107619.

Senior, Andrew W., Richard Evans, John Jumper, James Kirkpatrick, Laurent Sifre, Tim Green, Chongli Qin, Augustin Žídek, Alexander W. R. Nelson, Alex Bridgland, Hugo Penedones, Stig Petersen, Karen Simonyan, Steve Crossan, Pushmeet Kohli, David T. Jones, David Silver, Koray Kavukcuoglu, and Demis Hassabis. 2019. "Protein structure prediction using multiple deep neural networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13)." *Proteins: Structure, Function, and Bioinformatics* no. 87 (12):1141-1148. doi: https://doi.org/10.1002/prot.25834.

Sharma, Amit, Brij B. Gupta, Awadhesh Kumar Singh, and V. K. Saraswat. 2022. "Orchestration of APT malware evasive manoeuvers employed for eluding anti-virus and sandbox defense." *Computers & Security* no. 115:102627. doi: https://doi.org/10.1016/j.cose.2022.102627.

Wong, Chi Heem, Kien Wei Siah, and Andrew W Lo. 2018. "Estimation of clinical trial success rates and related parameters." *Biostatistics* no. 20 (2):273-286. doi: 10.1093/biostatistics/kxx069.

Yamazaki, Kashu, Viet-Khoa Vo-Ho, Darshan Bulsara, and Ngan Le. 2022. "Spiking Neural Networks and Their Applications: A Review." *Brain Sciences* no. 12 (7):863.

Yuhao Chen, Xin Tan, Borui Zhao, Zhaowei Chen, Renjie Song, Jiajun Liang, Xuequan Lu. 2023. "Boosting Semi-Supervised Learning by Exploiting All Unlabeled Data." *CVPR.*