ISSN: 2584-0355
DOI: doi.org/10.51219/MCCRJ/Julian-Lloyd-Bruce/406

URF PUBLISHERS
connect with research world

Medical & Clinical Case Reports Journal

https://urfpublishers.com/journal/case-reports

Vol: 3 & Iss: 4

Review

An Overview of Supervised Machine Learning in Drug Discovery, PK/PD Modeling and Precision Pharmacogenomics

Julian Lloyd Bruce, PhD*

Euclid University / Engelhardt School of Global Health and Bioethics, 1101 30th Street NW Suite #500 (Fifth Floor), Washington, D.C. 20007, USA

Citation: Bruce JL. An Overview of Supervised Machine Learning in Drug Discovery, PK/PD Modeling and Precision Pharmacogenomics. *Medi Clin Case Rep J* 2025;3(4):1439-1445. DOI: doi.org/10.51219/MCCRJ/Julian-Lloyd-Bruce/406

Received: 26 September, 2025; **Accepted:** 10 October, 2025; **Published:** 13 October, 2025

***Corresponding author:** Julian Lloyd Bruce, Euclid University / Engelhardt School of Global Health and Bioethics, 1101 30th Street NW Suite #500 (Fifth Floor), Washington, D.C. 20007, USA

ABSTRACT

Supervised machine learning (SML) is transforming pharmaceutical research by enabling precise, data-driven decision making across drug discovery, pharmacokinetics/pharmacodynamics (PK/PD) modeling, chemical synthesis and pharmacogenomics. This review synthesizes recent advances in SML applications across these domains and highlights how ensemble methods, graph-based architectures and hybrid mechanistic frameworks contribute to improved predictive accuracy, experimental efficiency and translational relevance. In drug discovery, SML accelerates virtual screening, predicts ADME properties and guides lead optimization. In PK/PD modeling, it supports individualized dose prediction, toxicity assessment and formulation design through the integration of multimodal clinical and molecular data. In chemical synthesis, SML improves reaction outcome prediction, retrosynthetic planning and condition optimization, enabling faster and more reliable route development. In pharmacogenomics, it advances genotype-informed dosing, adverse drug reaction prediction and treatment response modeling to support personalized medicine. Persistent challenges include data standardization, model interpretability, regulatory acceptance and ethical oversight. Overall, SML is a foundational technology with the potential to drive scalable, transparent and equitable innovation across the pharmaceutical landscape.

Keywords: Supervised machine learning; Artificial intelligence; Drug discovery; Pharmacogenomics; Pharmacokinetics (PK); Pharmacodynamics (PD); Chemical synthesis

Introduction

Artificial intelligence (AI) has become a transformative force in pharmaceutical research, led by supervised machine learning (SML) models that deliver precise, data-driven insights across the drug discovery and development pipeline. By training on labeled datasets to recognize patterns and make predictions, SML enables researchers to interrogate complex, high-dimensional data beyond the capabilities of traditional computational methods[1]. This approach is instrumental for optimizing compound screening, predicting therapeutic efficacy and refining dosing strategies with high accuracy[2].

The integration of SML into pharmaceutical research spans from early-stage drug discovery to clinical trials and personalized medicine. Algorithms such as quantitative structure-activity relationship (QSAR) models analyze structural and chemical properties to identify promising drug candidates early in development[3]. Virtual screening (VS) techniques, including advanced support vector machines (SVMs) and neural networks,

enable researchers to sift through expansive compound libraries with remarkable precision, expediting the drug discovery process.

Beyond discovery, SML's capacity to process multimodal datasets, encompassing genomics, imaging, chemical properties and patient histories, enhances clinical trial optimization and treatment personalization[4]. By leveraging genetic, molecular and clinical data, these models support tailored therapies that improve patient outcomes[5]. This capability addresses limitations that traditional computational methods struggle to overcome.

Despite its considerable impact, several challenges must be addressed for widespread clinical implementation. Model interpretability, data privacy concerns and computational demands remain key obstacles, prompting the exploration of novel approaches such as federated learning (FL) and hybrid machine learning (ML) frameworks[6]. Addressing these issues will be essential to unlocking SML's full potential in pharmaceutical sciences and healthcare.

As AI-driven methodologies continue to evolve, SML stands at the forefront of advancing precision medicine, shortening development timelines and driving innovation in healthcare. With ongoing advancements in automation and predictive analytics, SML is poised to reshape the future of medicine, ultimately leading to improved diagnostics, targeted treatments and enhanced patient care.

Methods

I ran a focused literature search on supervised machine learning (SML) in drug discovery, PK/PD modeling, chemical synthesis and retrosynthesis and pharmacogenomics. I searched five databases: PubMed, Google Scholar, Scopus, Web of Science and Embase. I used Medical Subject Headings (MeSH) and close keyword variants, including Machine Learning, Supervised, Classification, Regression, Support Vector Machines, Random Forest, Gradient Boosting, Graph Neural Networks, QSAR, Virtual Screening, Pharmacokinetics, Pharmacodynamics, Pharmacogenomics, Dose-Response Relationship, Treatment Outcome and Drug-Related Side Effects and Adverse Reactions. The main date range is 2019 to 2025, with earlier landmark papers added when needed for context.

I included peer-reviewed studies that clearly used supervised methods and reported enough detail to understand the data, the model, the training and validation approach and the metrics. I excluded opinion pieces, news items, unvalidated patents and studies without a baseline or a clear data split. After screening titles and abstracts, I read the full text of likely papers and kept those that met the criteria. I also checked references of included papers to find any important studies I missed.

Supervised machine learning: Foundations and methodology

SML is a subcategory of ML in which algorithms are trained on labeled datasets, where each input is paired with a known output. This structured training process enables models to learn from examples, capturing underlying relationships within data to make accurate predictions. By generalizing complex patterns, SML facilitates predictive modeling across diverse applications, particularly in classification and regression tasks. Classification tasks involve predicting categorical outcomes, such as diagnosing diseases from medical imaging or identifying fraudulent transactions based on behavioral patterns. Regression tasks focus on continuous predictions, such as estimating drug efficacy based on patient biomarkers or forecasting healthcare costs[7,8].

At its core, SML operates through an iterative optimization process aimed at minimizing an error function, commonly referred to as a loss function. The model is trained using input-output pairs, where each input (X) corresponds to a ground-truth output (Y). Through a series of computational steps, the model learns a function f(X) that maps X to Y while minimizing predictive error[7].

Key components of SML

- **Data preparation & Feature engineering:** The first step involves curating and preprocessing labeled datasets, ensuring data integrity, normalization and feature extraction. Feature selection is crucial in enhancing the model's ability to focus on relevant patterns while mitigating noise.
- **Model selection:** Depending on the problem type (classification or regression), different algorithms can be used. Linear regression models are common for continuous output predictions, whereas decision trees, support vector machines (SVMs) and neural networks excel at complex pattern recognition.
- **Training process:** The model iteratively adjusts internal parameters (weights) to minimize a loss function using techniques such as gradient descent. With each iteration, weights are updated to reduce the difference between predicted and actual outputs, thereby improving accuracy.
- **Performance evaluation:** Metrics such as mean squared error (MSE) for regression models and precision-recall, F1 score and accuracy for classification tasks help assess the model's effectiveness.
- **Deployment & Fine-tuning:** Once trained, the model is deployed and continuously refined through hyperparameter tuning and additional training on new data, ensuring long-term adaptability and performance stability in dynamic environments
- **Generalization & Overfitting prevention:** To ensure that the model performs well on unseen data, techniques such as L1/L2 regularization, dropout layers (for neural networks) and validation datasets are employed to prevent overfitting, where a model becomes too specialized to the training data[9,10].

Through these processes, SML enables robust decision making by leveraging structured data to develop predictive models that generalize effectively across real-world applications. Its broad applicability, ranging from medical diagnostics to financial forecasting, underscores its central role in modern AI-driven analytics.

SML methodologies in biotechnology and healthcare research

SML has become an essential tool in healthcare and pharmaceutical research, playing a vital role in classification and regression tasks that power diagnostic systems, drug efficacy modeling and personalized treatment strategies. To meet the specific demands of diverse medical datasets and clinical applications, a range of SML methodologies have been adapted and optimized accordingly:

- **Naïve Bayes (NB):** A probabilistic classifier based on Bayes' theorem that assumes feature independence, making it highly efficient for disease classification and genome analysis, particularly in handling high-dimensional genetic data.
- **K-Nearest Neighbors (KNN):** A nonparametric method that classifies data points based on proximity to labeled examples. KNN is commonly employed in patient stratification, anomaly detection and treatment recommendation systems.
- **Support Vector Machines (SVM):** By constructing optimal hyperplanes to separate classes in high-dimensional feature space, SVMs excel in complex tasks such as tumor classification and radiographic image interpretation, where subtle patterns must be discerned.
- **Ensemble Learning (Random Forest, Gradient Boosting):** These methods combine multiple weak learners to build more accurate and robust models. Ensemble techniques are frequently used in predictive diagnostics, disease risk modeling and biomarker selection.
- **Random Forest (RF):** As a specific ensemble method composed of decision trees, RF reduces overfitting and enhances reliability in both classification and regression. It is widely applied in pharmacogenomics, drug response prediction and multi-omics data integration.
- **Linear Regression (LiR):** A fundamental approach to modeling linear relationships between variables, LiR is heavily used in pharmacometrics to determine optimal dosing regimens and understand drug concentration-effect relationships.
- **Support Vector Regression (SVR):** A regression-specific variant of SVM that predicts continuous outcomes within a defined margin of tolerance. SVR is well suited to precision medicine applications, such as forecasting individualized treatment responses from genetic and molecular data[10-12].

The application of these SML methodologies enables effective generalization from large-scale biomedical datasets, reinforcing their indispensable role in drug discovery, diagnostics and treatment optimization. As computational power and data availability continue to grow, SML is poised to drive significant advancements in precision medicine, refining therapeutic strategies and improving patient outcomes.

Applications of SML in drug discovery and design

SML models have reshaped drug discovery and personalized medicine by improving the efficiency and accuracy of core workflows. A central advantage is the capacity to analyze and learn from large molecular datasets, which helps researchers rapidly identify compounds with promising therapeutic profiles. Techniques such as support vector machines (SVM), decision trees and random forest (RF) perform well for these tasks, using historical bioactivity data to predict efficacy, safety and bioavailability. For example, Korotcov, et al. reported that RF models outperformed deep neural networks in predicting the ADME properties of drug candidates across diverse chemical spaces, reinforcing the robustness of traditional SML approaches for early-stage screening[13].

In pharmacogenomics, SML has advanced personalized medicine by enabling precise dosing based on genetic and clinical features. Gradient-boosting methods such as CatBoost and XGBoost show strong performance in predicting warfarin maintenance doses when models include polymorphisms in genes like CYP2C9 and VKORC1, together with demographic and clinical variables. These models exceed the performance of linear regression by capturing nonlinear interactions and complex feature dependencies, which reduces adverse drug reactions and improves outcomes. By incorporating genomic variability, particularly variation in cytochrome P450 enzyme activity, these tools support a move away from generalized dosing toward adaptive, genotype-informed prescribing strategies[14,15].

Beyond screening and personalization, SML supports applications across pharmacokinetics (PK) and pharmacodynamics (PD). One recent study applied support vector regression (SVR) to predict methotrexate plasma concentrations in pediatric oncology, using individualized features such as age, body surface area, renal function and genetic polymorphisms. Compared with population-based PK models, SVR more accurately estimated peak and trough concentrations, captured nonlinear dose–exposure relationships without overfitting and improved safety in chemotherapy dosing. Such precision modeling enables tailored therapeutic windows and supports safer, more effective regimens in populations with high interindividual variability[16]. In cheminformatics and retrosynthesis, graph-based SML models, including Graph Neural Networks (GNNs) and message-passing neural networks, have been used to evaluate reaction feasibility and to predict synthesis routes, which reduces the time needed to identify viable pathways[17].

The practical impact of SML also includes economic and operational gains in drug development. As noted by Kumar, et al., SML can streamline early-stage screening by integrating chemical, biological and pharmacological data to prioritize candidates with higher probabilities of clinical success[18]. This data-driven strategy improves predictive accuracy, reduces reliance on costly trial-and-error methods and lowers the risk of late-stage failures. By focusing resources on high-potential leads, SML increases return on investment and shortens time to market. In parallel, precision-focused design supported by ML reduces adverse events and unnecessary interventions while optimizing patient outcomes and the use of healthcare resources.

The continued success of SML in drug discovery and personalized medicine depends on progress in areas such as integration with electronic health records (EHRs), data standardization, regulatory validation and clinician training for interpreting model outputs. As SML evolves across pharmacogenomics, PK and PD modeling and compound design, addressing these issues will be essential for translating computational advances into practical, scalable improvements in patient care. Overcoming these barriers will unlock the full potential of SML and accelerate the shift toward a data-driven, precision-oriented pharmaceutical ecosystem.

Pharmacokinetic and pharmacodynamic modeling

SML has emerged as a powerful framework for advancing pharmacokinetic (PK) and pharmacodynamic (PD) modeling. It offers a level of granularity and adaptability that traditional compartmental models often lack. By leveraging high-dimensional, multimodal datasets, SML enables more precise prediction of drug absorption, distribution, metabolism and elimination (ADME). These capabilities support individualized

dosing strategies, early toxicity screening and formulation optimization throughout the drug development process.

One foundational application of SML in PK modeling is the prediction of drug clearance and systemic exposure. Uno et al. (2024) demonstrated that random forest and support vector regression models, trained on clinical variables such as renal function, age and genetic polymorphisms, significantly outperformed conventional population PK models in predicting interindividual variability in drug clearance[19]. Their findings highlight the clinical utility of SML in early-phase trials, where accurate dose selection is essential for minimizing variability and optimizing therapeutic windows. Notably, their approach reduced residual error in clearance predictions, suggesting that SML can serve as a more reliable alternative to traditional covariate-based modeling for renally eliminated compounds.

This capacity for individualized modeling is especially impactful in pediatric oncology, where developmental pharmacology introduces substantial variability in drug metabolism. Tang, et al. applied SML to methotrexate and vincristine pharmacokinetics in children, incorporating demographic, clinical and laboratory features to predict plasma concentrations[20]. Their models achieved superior predictive accuracy compared to standard population-based approaches and enabled more precise dose adjustments. This reduced the risk of underexposure or toxicity and demonstrated how SML can overcome the limitations of one-size-fits-all dosing in vulnerable populations, where therapeutic margins are narrow and interpatient variability is high.

To balance model interpretability with predictive flexibility, Gharat, et al. proposed a hybrid modeling framework that integrates mechanistic PK/PD models with machine learning algorithms[21]. Their approach embeds physiological priors, such as enzyme kinetics and receptor occupancy, into data-driven models. This allows for both mechanistic insight and empirical adaptability. The hybrid framework showed improved generalizability across datasets and therapeutic classes, making it particularly valuable in complex disease areas like oncology and immunology. These fields often involve dynamic and partially understood biological systems. The integration of mechanistic and statistical modeling represents a promising direction for translational pharmacology, enabling models that are both explainable and responsive to real-world variability.

Beyond efficacy modeling, SML has proven instrumental in preclinical safety assessment. Chou, et al. used ensemble learning techniques, including gradient boosting and random forest, to predict drug-induced liver injury (DILI) based on chemical structure descriptors, transcriptomic data and in vitro assay results[22]. Their models identified early biomarkers of hepatotoxicity and stratified compounds by risk level with high sensitivity and specificity. This application shows how SML can function as a computational triage tool, reducing the likelihood of late-stage failures by flagging high-risk compounds early in development. Additionally, the integration of multi-omics data into predictive toxicology models reflects a broader trend toward systems-level modeling in drug safety.

In pharmaceutical formulation, SML has been applied to predict drug release kinetics from controlled-release systems under physiologically relevant conditions. Ota, et al. developed models that accurately forecasted both in vitro and in vivo dissolution profiles by training on formulation parameters, polymer characteristics and biorelevant media conditions[23]. Their work demonstrated that SML can reduce the need for iterative wet-lab testing and accelerate the optimization of extended-release formulations. This is especially valuable for complex dosage forms, where traditional empirical methods are time-consuming and resource-intensive.

Taken together, these studies illustrate the multifaceted role of SML in PK/PD modeling. SML is driving individualized dose optimization, enhancing hybrid mechanistic models, supporting early-stage toxicity assessments and guiding formulation design. These advances are being achieved with greater precision and efficiency than traditional approaches. As SML tools continue to evolve in interpretability, data efficiency and experimental validation, their integration into regulatory frameworks, clinical pharmacology and pharmaceutical engineering will be essential to realizing the full potential of data-driven precision medicine.

Chemical synthesis

SML is transforming chemical synthesis by enabling precise prediction of reaction outcomes, retrosynthetic pathways, optimal reaction conditions and selectivity profiles. Using extensive reaction databases and detailed molecular descriptors, SML models capture subtle structure-reactivity relationships that inform and refine synthetic planning. This data-driven approach reduces the need for exhaustive experimentation, accelerates discovery and expands access to complex molecular architectures, which makes synthesis more efficient and strategically guided by computational insight.

A key advance in this field was introduced by Coley, et al., who developed graph-convolutional neural networks (GCNNs) that represent molecules as graphs. This architecture allows the model to learn atom- and bond-level transformations directly from reaction data[24]. Their models achieved high accuracy in predicting major products across a wide range of reaction classes, outperforming rule-based expert systems and demonstrating the ability of SML to generalize beyond curated templates. Their work also emphasized the interpretability of learned chemical features, which enables chemists to trace predictions back to specific molecular substructures. This capability is essential for integrating AI into experimental workflows.

Building on this foundation, Strieth-Kalthoff, et al. reviewed SML applications in computer-aided synthesis planning. They highlighted how supervised models trained on reaction databases can identify viable disconnections and suggest plausible precursors for retrosynthetic analysis[25]. Their work marked a shift from rule-based retrosynthesis to data-driven route generation, where models learn from empirical precedent rather than manually encoded heuristics. This transition has broadened access to synthetic planning tools and has empowered chemists to explore novel pathways and scaffold modifications with greater speed and confidence.

Alnammi, et al. broadened the predictive scope of SML by incorporating reaction conditions, including temperature, solvent and catalyst, into models of yield and selectivity[26]. Their study showed that including contextual variables significantly enhances model performance, particularly in high-throughput experimentation where optimizing conditions is a major bottleneck. By combining chemical descriptors with experimental metadata, their framework accurately predicted

reaction outcomes under diverse conditions and offered a practical solution for guiding empirical screening while conserving resources.

Predicting selectivity, particularly regioselectivity and chemo selectivity, remains a major challenge in complex molecule synthesis. Zuranski, et al. addressed this challenge by training SML models on curated datasets of site-selective transformations. Their models captured subtle electronic and steric influences on reactivity, achieved high predictive accuracy and provided interpretable insights into the factors governing selectivity[27]. This work supports both mechanistic hypothesis generation and synthetic planning and it illustrates how SML can complement human intuition in navigating the multidimensional landscape of selectivity control.

To improve generalization and reduce overfitting, Oliveira, et al. introduced a multitask learning framework that predicts multiple reaction attributes, such as product identity, yield and reaction class, using shared molecular representations[28]. Their architecture leveraged interrelated chemical features across tasks, which enhanced model robustness and enabled more comprehensive reaction modeling. This multitask approach is especially valuable in low-data environments, where single-task models often struggle to capture nuanced reactivity patterns.

Recognizing the need for interpretability and uncertainty quantification, Rizvi Syed Aal E Ali, et al. proposed integrating attention mechanisms and confidence scoring into SML pipelines for reaction prediction[29]. Their study emphasized that actionable AI in chemistry must go beyond accuracy to provide transparent, confidence-calibrated outputs that chemists can rely on. By identifying which molecular substructures contributed most to a prediction and by quantifying uncertainty, their framework supports more informed decision making in both discovery and process chemistry.

Singh, et al. addressed data scarcity in reaction condition optimization by applying transfer learning and active learning strategies to SML models[30]. Their framework achieved strong predictive performance with limited experimental data and it showed that pre-trained models can be fine-tuned on small, domain-specific datasets to guide early-stage synthesis campaigns. This approach is particularly useful for rare or proprietary reaction classes, where large public datasets are not available.

Taken together, these advances show that SML is redefining chemical synthesis as a unified, end-to-end framework that includes forward reaction prediction, retrosynthetic design, condition optimization, selectivity modeling and uncertainty estimation. As SML models continue to improve in interpretability, data efficiency and experimental validation, they are poised to accelerate chemical discovery and expand the range of molecules that can be synthesized with precision and reliability.

Pharmacogenomics

Pharmacogenomics has progressed rapidly with the integration of SML, which enables the combined analysis of genomic, clinical and demographic data to predict individual drug responses and guide personalized treatment. By modeling complex, nonlinear interactions among genetic variants, SML algorithms support the transition from generalized, population-based dosing to truly individualized therapeutic strategies. This shift lays the foundation for more effective and precise frameworks in precision medicine.

One of the central challenges in pharmacogenomics is the high dimensionality and heterogeneity of genomic data, which often includes thousands of single nucleotide polymorphisms (SNPs) with modest effect sizes. Casale, et al. addressed this issue by applying SML algorithms to identify SNPs associated with variability in drug metabolism and response phenotypes across diverse populations[31]. Their study demonstrated that ensemble methods such as random forest and gradient boosting can effectively prioritize pharmacogenetically relevant variants while accounting for gene–gene and gene-environment interactions. This approach enhances both the interpretability and clinical utility of pharmacogenomic models, especially in multiethnic cohorts where allele frequencies and linkage disequilibrium patterns vary.

Cilluffo, et al. further explored the use of SML in predicting adverse drug reactions (ADRs) by integrating genomic and clinical data from pharmacovigilance databases[32]. Using support vector machines and decision tree classifiers, their models achieved high sensitivity and specificity in identifying patients at elevated risk for drug-induced hypersensitivity syndromes. Their work also emphasized the importance of feature selection and dimensionality reduction techniques, including recursive feature elimination and principal component analysis, which help mitigate overfitting and improve model generalizability. This study highlights the potential of SML to enhance drug safety by enabling the preemptive identification of individuals at risk based on genetic predisposition.

In psychiatric pharmacogenomics, Athreya, et al. developed a deep learning framework to predict antidepressant response in patients with major depressive disorder (MDD) using genomic and clinical features[33]. Their model, trained on data from the STAR*D (Sequenced Treatment Alternatives to Relieve Depression) trial, outperformed traditional statistical approaches in classifying responders and non-responders to selective serotonin reuptake inhibitors (SSRIs). To improve clinical applicability, the authors incorporated explainability techniques such as SHAP (SHapley Additive exPlanations), which helped identify key genetic markers and clinical variables driving model predictions. Integrating interpretability into deep learning pipelines is essential for clinical translation, as it allows clinicians to understand and trust model outputs when making therapeutic decisions.

Kalinin, et al. proposed a hybrid modeling approach that combines mechanistic pharmacogenomic knowledge with data-driven SML techniques to improve both prediction accuracy and biological plausibility[34]. Their framework integrates known gene–drug interaction networks with supervised learning models, allowing prior biological knowledge to inform the training process. This hybridization strengthens model robustness and interpretability, particularly in scenarios where training data are sparse or noisy. Their work illustrates the value of embedding domain expertise into machine learning pipelines to bridge the gap between computational prediction and clinical relevance.

Finally, Tafazoli, et al. demonstrated the utility of SML in predicting warfarin dose requirements based on genetic polymorphisms in CYP2C9, VKORC1 and CYP4F2, along with

demographic and clinical variables[35]. Their study compared multiple SML algorithms, including random forest, support vector regression and artificial neural networks and found that ensemble models yielded the most accurate dose predictions across diverse patient populations. This research reinforces the role of SML in refining pharmacogenetic dosing algorithms, especially for drugs with narrow therapeutic indices and high interindividual variability.

Taken together, these studies illustrate how SML is reshaping pharmacogenomics by converting complex, multidimensional datasets into clinically actionable insights. Whether predicting adverse drug reactions, modeling antidepressant response, refining warfarin dosing or integrating domain knowledge into hybrid frameworks, SML provides a scalable and interpretable pathway toward truly personalized drug therapy. This approach anchors treatment decisions in the rich context of each patient's genetic profile.

Limitations and Challenges

Although SML holds transformative potential for pharmaceutical research, several limitations continue to impede its widespread adoption in clinical and industrial settings. These challenges include issues related to data quality, model interpretability, regulatory integration and ethical considerations. Each of these must be addressed to fully unlock the promise of AI-driven drug development.

One persistent barrier is the lack of high-quality, standardized datasets. Mathrani, et al. emphasize that biomedical data often suffer from heterogeneity, including inconsistent labeling, missing values and variable measurement protocols[36]. Such inconsistencies undermine model generalizability and reproducibility across institutions and populations. This problem is especially pronounced in multi-center studies, where differences in data collection and annotation can introduce bias and reduce the external validity of trained models. Overcoming this challenge will require the establishment of harmonized data standards and the development of robust preprocessing pipelines capable of accommodating real-world variability without compromising model performance.

From a regulatory perspective, Yang, et al. argue that the integration of SML into clinical workflows is constrained by the absence of standardized validation frameworks and clear guidelines for model approval[38]. Unlike traditional statistical models, SML algorithms often evolve over time through retraining and fine-tuning, raising questions about version control, auditability and long-term reliability. Regulatory bodies such as the FDA and EMA are beginning to address these issues, but a consensus on best practices for model validation, monitoring and lifecycle management is still emerging.

Ethical and equity considerations also pose significant challenges. Obaido, et al. underscore the risk of algorithmic bias, particularly when models are trained on datasets that underrepresented minority populations or reflect historical inequities in healthcare access[39]. Such biases can propagate through predictive pipelines, leading to disparities in treatment recommendations and outcomes. Ensuring fairness in SML requires proactive bias auditing, inclusive data collection and the implementation of fairness-aware learning algorithms that explicitly account for demographic variability.

In summary, although SML holds immense promise for advancing drug discovery and tailoring treatments to individual patients, its real-world impact hinges on addressing persistent challenges in data reliability, model transparency, regulatory compliance and ethical oversight. Tackling these barriers is critical to developing AI systems in pharmaceutical science that are not only effective, but also trustworthy, equitable and clinically meaningful.

Conclusion

SML is redefining pharmaceutical research by enabling scalable, data-driven approaches to drug discovery, development and precision medicine. Its ability to integrate and model complex biological, chemical and clinical data has accelerated key processes such as compound screening, pharmacogenomic profiling, PK/PD modeling and chemical synthesis. From improving warfarin dosing accuracy to predicting reaction outcomes and adverse drug events, SML tools now inform therapeutic decision making with a level of precision that surpasses traditional statistical methods. Emerging techniques such as ensemble learning, graph-based models and hybrid mechanistic frameworks have expanded both the interpretability and performance of SML systems, making them increasingly relevant across clinical and industrial contexts.

Yet the successful integration of SML into real-world pharmaceutical workflows requires overcoming persistent challenges related to data heterogeneity, model transparency, regulatory validation and ethical accountability. Addressing these issues is critical to ensuring that SML systems are not only predictive but also trustworthy, equitable and aligned with the standards of clinical care. As interdisciplinary collaboration deepens and regulatory frameworks evolve, supervised learning is poised to become a cornerstone of next-generation drug development. It holds the potential to accelerate discovery, personalize therapy and improve patient outcomes across diverse populations.

Declarations

This literature review did not involve human or animal subjects; therefore, ethics approval and consent to participate were not required. No personal details, images or videos of individuals are included in the manuscript and consent for publication is not applicable.

All data and materials referenced are publicly available or cited appropriately; no proprietary datasets were used. The author declares no competing interests and no external funding was received to support this research.

References

1. Vora LK, Gholap AD, Jetha K, Thakur RR, Solanki HK, Chavda VP. Artificial intelligence in pharmaceutical technology and drug delivery design. Pharmaceutics 2023;15(7):1916.
2. Hutson M. How AI is being used to accelerate clinical trials. Nature 2024;627(8003):2-5.
3. Qi X, Zhao Y, Qi Z, Hou S, Chen J. Machine learning empowering drug discovery: Applications, opportunities and challenges. Molecules 2024;29(4):903.
4. Askr H, Elgeldawi E, Aboul Ella H, Elshaier YA, Gomaa MM, Hassanien AE. Deep learning in drug discovery: an integrative review and future challenges. Artificial Intelligence Review 2023;56(7):5975-6037.

5. Suriyaamporn P, Pamornpathomkul B, Patrojanasophon P, Ngawhirunpat T, Rojanarata T, Opanasopit P. The artificial intelligence-powered new era in pharmaceutical research and development: a review. AAPS PharmSciTech 2024;25(6):188.
6. Bhatia N, Khan MM, Arora S. The Role of Artificial Intelligence in Revolutionizing Pharmacological Research. Current Pharmacology Reports 2024;10(6):323-329.
7. Osisanwo FY, Akinsola JE, Awodele O, et al. Supervised machine learning algorithms: classification and comparison. Int J Computer Trends Tech (IJCTT) 2017;48(3):128-138.
8. Uddin S, Khan A, Hossain ME, Moni MA. Comparing different supervised machine learning algorithms for disease prediction. BMC medical informatics and decision making 2019;19(1):1-6.
9. Shetty SH, Shetty S, Singh C, Rao A. Supervised machine learning: algorithms and applications. Fundamentals and methods of machine and deep learning: algorithms, tools and applications 2022:1-6.
10. Choudhary R, Gianey HK. Comprehensive review on supervised machine learning algorithms. In 2017 Int Conf Machine Learning Data Sci (MLDS) 2017:37-43.
11. Abdel-Jaber H, Devassy D, Al Salam A, Hidaytallah L, El-Amir M. A review of deep learning algorithms and their applications in healthcare. Algorithms 2022;15(2):71.
12. Shailaja K, Seetharamulu B, Jabbar MA. Machine learning in healthcare: A review. In2018 Second international conference on electronics, communication and aerospace technology (ICECA) 2018:910-914.
13. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. Molecular pharmaceutics 2017;14(12):4462-75.
14. Raja K, Patrick M, Elder JT, Tsoi LC. Machine learning workflow to enhance predictions of Adverse Drug Reactions (ADRs) through drug-gene interactions: application to drugs for cutaneous diseases. Scientific reports 2017;7(1):3690.
15. He J, Wu Y, Yuan L, et al. An inductive learning-based method for predicting drug-gene interactions using a multi-relational drug-disease-gene graph. J Pharmaceutical Analysis 2025:101347.
16. Jian C, Chen S, Wang Z, et al. Predicting delayed methotrexate elimination in pediatric acute lymphoblastic leukemia patients: an innovative web-based machine learning tool developed through a multicenter, retrospective analysis. BMC Medical Informatics and Decision Making. 2023;23(1):148.
17. Schwaller P, Gaudin T, Lanyi D, et al. Found in Translation: predicting outcomes of complex organic chemistry reactions using neural sequence-to-sequence models. Chemical science 2018;9(28):6091-6098.
18. Kumar SA, Ananda Kumar TD, Beeraka NM, et al. Machine learning and deep learning in data-driven decision making of drug discovery and challenges in high-quality data acquisition in the pharmaceutical industry. Future Med Chem 2022;14(4):245-270.
19. Uno M, Nakamaru Y, Yamashita F. Application of machine learning techniques in population pharmacokinetics/pharmacodynamics modeling. Drug Metabolism and Pharmacokinetics 2024;56:101004.
20. Tang A. Machine learning for pharmacokinetic/pharmacodynamic modeling. J Pharmaceutical Sci 2023;112(5):1460-1475.
21. Gharat SA, Momin MM, Khan T. Artificial Intelligence and Machine Learning in Pharmacokinetics and Pharmacodynamic Studies. In Pharmacokinetics and Pharmacodynamics of Novel Drug Delivery Systems: From Basic Concepts to Applications: A Machine-Generated Literature Overview 2024:343-393.
22. Chou WC, Lin Z. Machine learning and artificial intelligence in physiologically based pharmacokinetic modeling. Toxicological Sciences 2023;191(1):1-4.
23. Ota R, Yamashita F. Application of machine learning techniques to the analysis and prediction of drug pharmacokinetics. J Controlled Release 2022;352:961-969.
24. Coley CW, Green WH, Jensen KF. Machine learning in computer-aided synthesis planning. Accounts of chemical research 2018;51(5):1281-1289.
25. Strieth-Kalthoff F, Sandfort F, Segler MH, Glorius F. Machine learning the ropes: principles, applications and directions in synthetic chemistry. Chemical Society Reviews 2020;49(17):6154-6168.
26. Alnammi M, Liu S, Ericksen SS, et al. Evaluating scalable supervised learning for synthesize-on-demand chemical libraries. J Chem Information Modeling 2023;63(17):5513-5528.
27. Zuranski AM, Martinez Alvarado JI, Shields BJ, Doyle AG. Predicting reaction yields via supervised learning. Accounts of Chem Res 2021;54(8):1856-1865.
28. Oliveira JC, Frey J, Zhang SQ, et al. When machine learning meets molecular synthesis. Trends in Chemistry 2022;4(10):863-885.
29. Ali RS, Meng J, Khan ME, Jiang X. Machine learning advancements in organic synthesis: A focused exploration of artificial intelligence applications in chemistry. Artificial Intelligence Chemistry 2024;2(1):100049.
30. Singh S, Sunoj RB. Molecular machine learning for chemical catalysis: prospects and challenges. Accounts of Chemical Research 2023;56(3):402-412.
31. Casale AD, Sarli G, Bargagna P, et al. Machine learning and pharmacogenomics at the time of precision psychiatry. Current Neuropharmacology 2023;21(12):2395-408.
32. Cilluffo G, Fasola S, Ferrante G, et al. Machine learning: An overview and applications in pharmacogenetics. Genes 2021;12(10):1511.
33. Athreya AP, Neavin D, Carrillo-Roa T, et al. Pharmacogenomics-driven prediction of antidepressant treatment outcomes: a machine-learning approach with multi-trial replication. Clin Pharmacology Therapeutics 2019;106(4):855-65.
34. Kalinin AA, Higgins GA, Reamaroon N, et al. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics 2018;19(7):629-650.
35. Tafazoli A, Mikros J, Khaghani F, et al. Pharmacovariome scanning using whole pharmacogene resequencing coupled with deep computational analysis and machine learning for clinical pharmacogenomics. Human Genomics 2023;17(1):62.
36. Mathrani A, Susnjak T, Ramaswami G, Barczak A. Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. Computers and Education Open 2021;2:100060.
37. Dinsdale NK, Bluemke E, Sundaresan V, et al. Challenges for machine learning in clinical translation of big data imaging studies. Neuron 2022;110(23):3866-3881.
38. Yang HS, Rhoads DD, Sepulveda J, et al. Building the model: challenges and considerations of developing and implementing machine learning tools for clinical laboratory medicine practice. Archives of patholog laboratory medicine 2023;147(7):826-836.
39. Obaido G, Mienye ID, Egbelowo OF, et al. Supervised machine learning in drug discovery and development: Algorithms, applications, challenges and prospects. Machine Learning with Applications 2024;17:100576.