Advanced International Journal of Multidisciplinary Research
E-ISSN: 2584-0487 editor@aijmr.com
Volume 2, Issue 5, September-October 2024
CrossRef DOI: 10.62127/aijmr.2024.v02i05.1077

Precision Medicine with Data-Driven Approaches: A Framework for Clinical Translation

Simranjit Kaur[1], Rowan Kim[2], Nisha Javagal[3], Joseph Calderon[4], Senia Rodriguez[5], Nithin Murugan[6], Kelsang Gyatsho Bhutia[7], Karan Dhingra[8], Saloni Verma[9]

[1]Independent Researcher, British Columbia, Canada
[2]Independent Researcher, Washington, United States
[3]Independent Researcher, Tennessee, United States
[4]Independent Researcher, Texas, United States
[5]Independent Researcher, Peru
[6]Independent Researcher, Illinois, United States
[7]Department of Medicine, Tashkent Medical Academy, Tashkent, Uzbekistan
[8]Department of Biomedical Engineering, University of Ottawa, Ontario, Canada
[9]Department of Biomedical Engineering, Cornell University, New York, United States

Abstract

Precision medicine-based approaches differentiate themselves by taking into consideration subpopulation variability (e.g. genetic variations, age, gender, race, addictions). To date, traditional models, such as Trial-and-Error Dosing, Empirical Treatment Guidelines, Statistical and Actuarial Models, Pathophysiological Models, and Clinical Judgment and Experience, have been generalized in healthcare fields. However, more comprehensive and innovative technologies are required besides these conventional modelings, which are subject to various limitations such as low efficiency and incapability of processing complex biological systems. Here we review diverse machine learning (ML) algorithms integrated with big data and omics and its applications in various aspects of precision medicine. ML is the branch of artificial intelligence (AI), which has been rapidly developed and highlighted as a promising method to decrease diagnostic errors and aid clinicians with decision-making in recent decades. We focused on applications of ML models such as support vector machine (SVM), K Nearest Neighbor (KNN) random forest (RF), convolutional neural networks (CNNs) and deep learning in drug toxicity prediction, cardiovascular diseases, neurodegenerative diseases, and cancer therapies within precision medicine and specific benefits and challenges of each. This review provides insights on the wider utilization in clinical environments by recognizing current advantages that are expected to expand the scope of AI-driven methods and issues that need to be addressed for further studies.

Keywords: Precision Medicine, Biosensors, Artificial Intelligence, Cancer Therapy

1. Introduction

The concept of enhancing healthcare fields by tailoring treatment to individuals depending on their specific characteristics is centuries old and remains a key component of medical practice. Awareness that patient heterogeneity was significant in treatment assessments began to emerge in the late twentieth century among both clinicians [1] and biostatisticians [2]. This patient heterogeneity implied the need of individualized therapy based on evidence-based medicine stated by Kraviz et al. [3]. These component ideas were combined and yielded the modern concept of precision medicine paradigm in which patient heterogeneity is leveraged via data-driven approaches in order to improve treatment decision-making so that the certain therapy is prescribed to the certain patient at the right time. Precision medicine became considered significant with President Obama's announcement of the Precision Medicine Initiative in this 2015 State of the Union Address [4].

In terms of extracting 'precise', as the name suggests, population-level data related to particular factors to the risk of disease, computational methods such as artificial intelligence (AI) and machine learning have stood out in the past decades. By the late 1700s the statistical measures had impacted, but actual advances started to dramatically increase after the first arbitrarily controlled clinical trial was successfully performed by Austin Bradford Hill in 1946 [5]. By the 1960s, statistical measures had been developed that enabled claims that specific biomarkers, behaviors, and other individual patient data could make the development of heart disease [6]. In the 1950s, McCarthy et al. [7] proposed AI as a prediction machine and Samuel [8] developed machine learning in 1959, leading to the investigation of discrimination of cells in microscopic images with machine learning at that time. Dechter [9] subsequently proposed deep learning in 1986 and Lecun et al. [10] proposed a convolutional neural network in 1988.

Moreover, AI-based applications for follow-up of treatment progress are also being actively developed and 13 applications were approved by the FDA from 2017 to 2020 [11]. The paradigm of medicine has been shifted from generalized solutions for the largest number of patients toward prevention, personalization, and precision with the development of AI technologies [12]. According to a recent National Academy of Medicine report about current and future state of AI in healthcare, they noted "unprecedented opportunities" as to augmentation of the specialists care and the AI-provided assistance in tackling the realities of being human (e.g. fatigue and inattention) and the risks of machine error [13]. With these verified potential of AI-based technologies, we delved deeply into the applications of AI in precision medicine including drug toxicity medicine, cardiovascular diseases, neurodegenerative diseases, and cancer therapies. Moreover, this review includes several challenges that we're currently encountering.

Precision medicine uses tools such as omics, pharmaco-omics, Big Data, Artificial Intelligence (AI), and machine learning (ML) to determine the best course of action when developing cures for patients [14]. Therefore, it has shown itself to be greatly relied upon in the evolution of healthcare, especially in clinical settings. One reason for its success is the use of computational methods. ML helps to recognize patterns in complicated datasets, allowing for a more effective organization of data. It does this by using a punishment and reward system in supervised models which instructs the AI to repeat good behaviors. This helps in drug toxicity prediction, an area in which ML shines [15]. Due to its thoroughness, ML combats challenges that might arise from precision medicine, such as large amounts of data, by carefully sorting through information and categorizing it. Like ML, computational modeling techniques also use number based data to enhance precision medicine [16]. They excel at grouping patients into similar sub-populations and finding data similarities between the individuals in the groups. Similarly, they help aid clinicians in determining treatments, making precision medicine safer for implementation into clinical settings. To retain credibility, another challenge brought up by the implementation of computational techniques, they follow a ten step checklist. Although precision medicine is a modern and valuable approach to medicine, it still faces challenges when implementing computational techniques into clinical practice methods. In this review, we show utilization of data-driven approaches (such as ML, deep learning algorithms, etc.) in various scopes of precision medicine (drug discovery, cardiovascular diseases, molecular biology diseases, etc.) to investigate the challenges of introducing computational techniques to clinical practice and possible solutions.

2. Discussion

A. Need for Modeling in Healthcare

Computational modeling has enhanced healthcare. It utilizes mathematical simulations and algorithms to advance precision medicine and clinical decision making, and it has had a notable impact on modern medical practices, specifically because it allows for increased personalization and complexity handling. Traditional approaches for analyzing healthcare have many limitations, increasing inefficiency, and risks. However, the use of modeling to counter these limitations provides many benefits such as those listed below [16].

Personalized Medicine and Patient Stratification

Personalized medicine and patient grouping leverage analysis of complex and heterogeneous patient data. Identifying subgroups with similar characteristics and allowing targeted treatments and strategies [17]. This can optimize resource allocation; additionally, it can improve patient outcomes through organization.

Clinical Decision Support and Treatment Optimization

Incorporating patient data such as genetic profiles or medical history provides clinical decision support. This approach offers personalized treatment recommendations for healthcare professionals and physicians to utilize [18].

Understanding Biological Concepts

Computational models can display the usage of data sources (such as genomics, proteomics, etc.) to provide scientific knowledge [17]. As a result, this can help create targeted treatments to advance precision medicine and healthcare abilities.

Drug Development and Safety Evaluation

The drug development process does involve predicting drug side effects and optimizing drug design. By simulating interactions between drugs and biological systems, safety concerns can be identified to increase safety and effectiveness for the therapeutic agents; this reduces the risk of adverse events, benefiting patients [18].

Resource Optimization and Healthcare Efficiency

Simulations and analysis of complex data (purposes of computational modeling) can improve resource allocation [19]. Overall results can include efficient utilization of resources, reduced costs, improved patient flow, and better quality and accessibility of healthcare services.

B. Data-driven predictions using machine learning techniques

Machine learning is a computer science paradigm which improves the ability to identify complex patterns in large datasets. It has been further classed as supervised, unsupervised, or reinforcement models. Supervised machine learning algorithms utilize labeled data to identify patterns in multidimensional data, such as identifying healthy and diseased people or predicting result scores. Punishments and rewards instruct reinforcement models to repeat good decisions (reward them) and avoid making bad decisions (punish them) in the future. A training data set with known labels is used to create a model and optimize its performance for the desired result [20]. Patterns discovered in the data can be used to classify new data sets or create individualized predictions. Models used for classification can group data into classes, while regression models can generally predict continuous outcomes [21].

The first key stage in drug discovery is to identify significant compounds. We now have access to a variety of biomedical databases to assist us in accomplishing this goal. In this context, target identification uses gene expression to better understand disease mechanisms and identify genes that contribute to the disease. The use of microarray and RNA-seq technologies has resulted in large amounts of data on gene expression in a variety of disorders [22]. Through the analysis of gene expression signatures, it is possible to identify the target genes associated with different disorders. For instance, van IJzendoorn et al. (2019) employed a machine learning approach and gene expression data to discover new biomarkers and potential drug targets for rare soft tissue sarcoma. In recent times, numerous computer programming and software have been developed to utilize different algorithms for interpreting results using predefined scoring functions. However, this remains a challenging task because energy levels and force fields for screening potential therapeutic compounds are extremely difficult to predict with certainty. Quantum physics has the potential to significantly improve efficiency in predicting future medication discoveries and reducing errors [15].

Due to the paradigm shift in genomics, new machine learning algorithms are applied to precision medicine in some areas. The K Nearest Neighbor algorithm is famous for its simplicity, widespread use, and high efficiency for pattern recognition. It merely classifies samples based on the category of their nearest neighbor. This classifier is straightforward, but it can be affected by redundant and irrelevant attributes when dealing with large volumes of medical data. Due to the elastic network's nature, oriented feature selection linear SVM can provide automatic feature selection, and the fused lasso ensures the smoothness of the coefficient vector. Compared to the approach of classifiers, where feature selection is an independent step, this is time-saving but brings about its own problems. Although features are selected, they may not provide the reliability of forward- and backward-stepwise selection [15]. Support vector machines (SVMs) are well-known machine learning algorithms for classification and regression applications. In healthcare, they have been used for a variety of tasks, including diagnosis, prognosis, and prediction of disease outcomes. SVMs are adept at addressing complex medical data due to their ability to analyze and address nonlinear relationships between features and classes. The ability to identify an optimal decision boundary representing the largest separation between classes is the key strength of an SVM. The creation of the optimal hyperplane is influenced by a small subset of training samples, which are the pivotal data structure in an SVM. The optimal hyperplane is not affected by the removal of the training samples that are not relevant to the SVs. Initially, SVMs were used to tackle linearly separable problems, with their capabilities later extended to handle nonlinear ones. Samples are mapped from a finite-dimensional space to a higher-dimensional space [23]. Random forests are models that are used for both classification and regression tasks. In a random forest a collection of decision trees are created, each trained in a different subset with some randomness utilized during the building process. Random Forests are part of the family of ensemble learning methods where multiple models are combined to improve accuracy and reduce overfitting [24].

C. Applications in Precision Medicine

Machine learning methods in drug toxicity prediction

Machine learning and deep learning algorithms have emerged as a method for predicting aspects of drug toxicity. These methods offer insights into the risks linked to pharmaceutical drugs, such, as carcinogenicity, acute oral toxicity and long term toxicity forecasts [15]. Toxicity is classified into two categories; acute and chronic. Examples of chronic toxicity include mutagenicity, carcinogenicity, and acute oral toxicity. Lethal Concentration 50, or LC50, is utilized in environmental studies, while Lethal Dose 50, or LD50, is frequently employed in research to assess toxicity. Because it can reduce the costs and personnel required for pharmaceutical clinical trials, toxicity prediction is crucial. Predicting toxicity accurately can help avoid a great deal of pharmacological evaluations, such as animal, cell, and clinical studies. The subject of toxicity prediction can benefit considerably from the incorporation of machine learning in the era of Big Data and artificial intelligence. In the context of Big Data and artificial intelligence, machine learning has great potential to assist with toxicity prediction since it can improve prediction accuracy by combining chemical structure descriptions with research of human transcriptome data. When paired with transcriptome data analysis, machine learning techniques such as support vector machines, random forests, deep learning, and k-nearest neighbors have demonstrated exceptional performance in toxicity prediction, improving the accuracy of toxicity prediction models. By using machine learning methods, abstract chemical fragments can be created and then used to create high-performance toxicity prediction models that are based on large amounts of data [25].

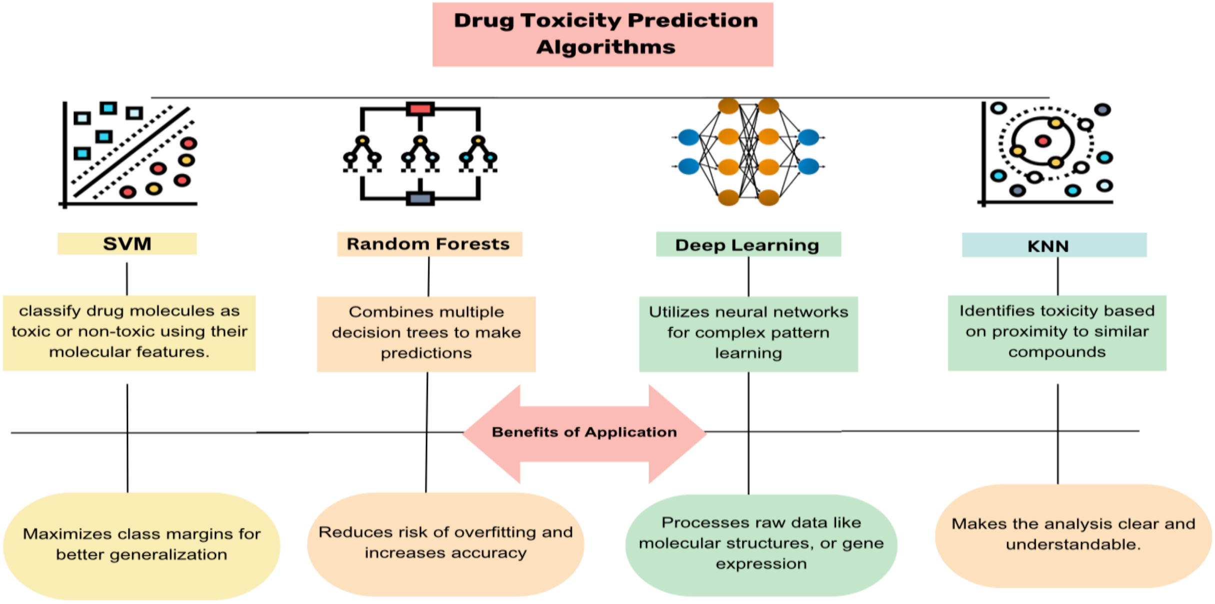
Drug Toxicity Prediction Algorithms
SVM
Random Forests
Deep Learning
KNN
classify drug molecules as toxic or non-toxic using their molecular features.
Combines multiple decision trees to make predictions
Utilizes neural networks for complex pattern learning
Identifies toxicity based on proximity to similar compounds
Benefits of Application
Maximizes class margins for better generalization
Reduces risk of overfitting and increases accuracy
Processes raw data like molecular structures, or gene expression
Makes the analysis clear and understandable.

Figure 1. Drug Toxicity Prediction Algorithms

To classify drug compounds as toxic or non-toxic, numerous modern drug toxicity prediction algorithms are used, each with its own set of strengths and methodologies. These algorithms combine to make drug toxicity assessments clearer and more understandable, each adding a layer of resilience and precision to the prediction models used in pharmacological research. Carcinogenicity is important in the development of drugs because of the serious health implications that can come from exposure to carcinogens. To be able to predict the carcinogenic potential of drugs, these models are trained by machine learning algorithms using information from animal studies and molecular descriptors. Therefore, such models enable early identification of chemicals that include cancer causing substances, which helps improve the decision-making process on medication development and reduce possible negative consequences on health. Mutagenicity is linked to increased risk of cancer and other genetic disorders as it is caused by substances that have the ability to change genetic material. This is why mutagenic power prediction models usually rely on data derived from tests like Ames test that examines mutagenic powers of compounds using bacterial strains. Whether a substance is mutagenic can be determined through machine learning methods that use molecular descriptors and empirical data. These models are helpful in using safer drug design and development processes by evaluating a large number of chemicals for mutagenicity. Hepatotoxicity, often known as drug-induced liver injury, or DILI, is a major issue in clinical practice and pharmaceutical research. Liver function tests reveal that harmful compounds can cause anything from mild liver problems to severe liver failure. Machine learning techniques are used to predict hepatotoxicity by analyzing chemical structures and other molecular parameters. These prediction techniques reduce the possibility of liver-related side effects by accurately identifying substances that potentially cause liver toxicity and helping to prioritize treatment alternatives. Determining safe dosages and possible health risks from exposure depends on evaluating the acute oral toxicity of substances. Pharmacological toxicity during oral administration is predicted by machine learning techniques utilizing molecular descriptors and data from animal research. These models are essential for ensuring patient safety, directing dose selection, and identifying high-toxicity compounds early in the pharmaceutical development process. Since hERG (Human Ether-a-go-go-Related Gene) is linked to QT prolongation, a potentially fatal heart rhythm problem, blocking it throughout the pharmaceutical development process is challenging. It is important to be able to predict hERG inhibition, which is needed to ascertain the presence of possible drug-induced cardiovascular abnormalities that require more caution. Therefore, in this case, safety pharmacology provides insights into potential mechanisms through which various drugs exert their toxicities. These predictive models are useful tools for analyzing compounds for possible heart toxicity, ultimately leading to the development of safer medications with fewer cardiovascular adverse effects [15].

Machine learning algorithms such as support vector machines (SVM) and k-nearest neighbors (k-NN) have been shown to improve standard QSAR models for predicting drug toxicity. It has been demonstrated that by optimizing traditional QSAR models with biological algorithms, random forests, and artificial neural networks, machine learning techniques like SVM and k-NN can improve their ability to forecast drug toxicity. Performance is greatly impacted by these methods, which increase prediction accuracy by taking datasets and computational representations into consideration. To be more precise, k-NN has proven to be more successful in predicting the oral LOAEL of rats. It has achieved high AUC values of up to 0.814, indicating its efficacy in toxicity prediction. The effectiveness of SVM in toxicity prediction models is demonstrated by the fact that it has been used to classify substances based on toxicity, with an average AUC value of 0.91. By integrating these methods with other molecular descriptors like Dragon descriptors, PubChem keys, and MACCS fingerprints, the accuracy of toxicity prediction is considerably increased by utilizing Big Data and artificial intelligence. Furthermore, the combination of human transcriptome data analysis with chemical structure descriptions constitutes a considerable leap in machine learning approaches, resulting in a large increase in prediction accuracy. This shift towards including transcriptome data allows for a more comprehensive understanding of biological mechanisms and cellular responses to harmful chemicals, resulting in more accurate drug toxicity predictions [25].

Big Data approaches to Cardiovascular medicine

Big data holds immense potential for cardiovascular medicine, yet managing and interpreting these large-scale datasets poses a significant challenge. It includes a broad spectrum of data types, including electronic health records (EHRs), social media data, genetics, and metabolomics. Combining these databases could lead to revolutionary changes in clinical trials, personalized medicines, and precision cardiovascular medicine. Chronic and varied cardiovascular diseases (CVDs) have been found and generally classified into phenotypes based on how they appear clinically. But given the complexity of chronic CVDs, it's possible that several causes can result in symptoms that are identical. This implies that while typical therapies are based on broad descriptions of the condition, individual responses to

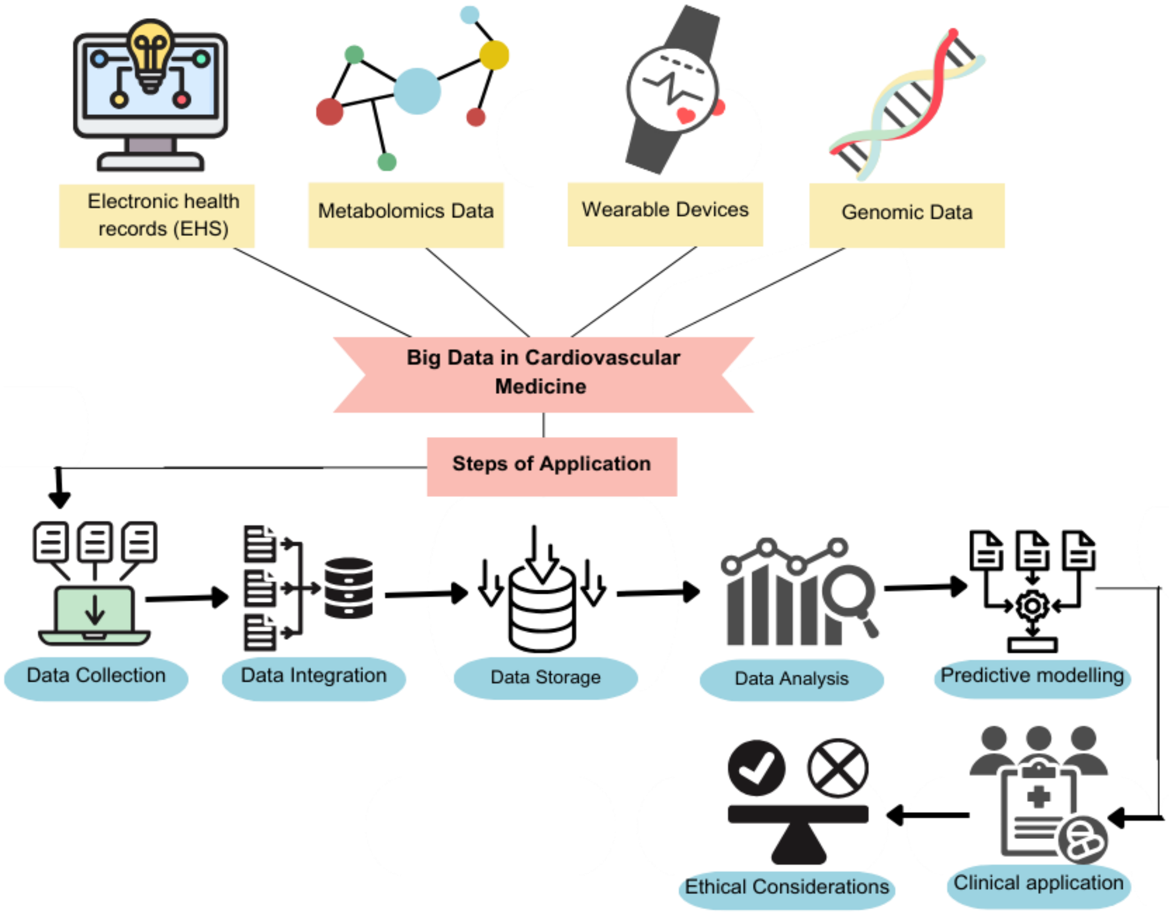

Figure 2. Big Data in Cardiovascular Medicine

them may vary. The environment, lifestyle choices, metabolism, and heredity all have an impact on CVDs. Due to their various causes and effects, different cardiovascular conditions require different treatment approaches [26]. Utilizing big data in cardiovascular medicine is a multifaceted process that involves numerous forms of data and processes to properly apply this data for improved treatment results. However, in order to respect patient rights and use big data to improve cardiovascular health, these improvements are susceptible to ethical considerations.

A chronic autoimmune disease that damages and inflames joints is called rheumatoid arthritis (RA). Primary signs of RA are joint related and RA patients are susceptible to stroke and CVD (covering heart attacks, strokes, and atherosclerosis). Risk factors for CVD and stroke in RA patients are smoking, obesity, high blood pressure, metabolic syndrome, and abnormal lipid levels. The AI techniques can help to devise more accurate risk assessment models by continuous learning from the data resulting in better prediction of CVD and stroke risk in RA patients. Combination of different biomarkers like OBBM, LBBM, RBBM and GBBM through the AI techniques will also help to better evaluate CVD and stroke risk in RA patients. The latest advancement in the field speculates to allow the use of combined genetic and AI platforms to check the severity of CVD and stroke in RA patients [27]. Shameer et al (2018) used the genetic data and deep learning algorithms on the electronic health records to devise a predictive model for the unfavorable cardiac events. The study explained to us how the AI-based risk score models may help to identify the people who are at high risks of heart problems and proper interventions can be made with special preventive measures. AI may alter the future of cardiac imaging and treatment through AI-driven precision healthcare. Convolutional neural networks (CNNs) were trained on cardiac imaging data in a different study by Wolterink et al. (2019) to automatically segment and measure cardiac regions such the left ventricle and coronary arteries. The study demonstrated how AI-assisted image processing could improve diagnostic precision, improve the interpretation of cardiac imaging tests, and support patients with heart problems in making treatment decisions [28]. The majority of individuals believe cardiovascular diseases (CVDs) are inherited. A lower risk of coronary heart disease is linked to protective genetic variants, like protein-inactivating variants in the NPC1L1 gene. On the other hand, specific gene mutations such as those in the LDLR gene indicate familial hypercholesterolemia.

The limitations of large-scale Genome-Wide Association Studies (GWAS) lie in their inability to identify pathogenic genes due to statistical power constraints and variations in genetic diversity among populations. An alternative approach would be focusing on genes associated with stress adaptation or resilience— which can provide protection against adverse cardiovascular phenotypes. Through big data, genomics can be better related to precision phenotyping in cardiovascular medicine; a move that sees increased sample sizes coupled with a sophisticated characterization of nuanced pathophenotypes, alongside ideal analytical strategies like network medicine. Transcriptomics, proteomics, and metabolomics are examples of multi-dimensional data that network medicine integrates with clinical data to understand complicated cardiovascular illnesses at numerous biological levels. Precision medicine therapies targeted at restoring normal network function and improving clinical outcomes in individuals with cardiovascular disorders are made possible by network medicine through the analysis of personalized disease modules. Network medicine leverages unsupervised analyses of proteomic data to identify patient subgroups and refine biological classifications, offering potential for personalized treatment strategies. It introduces the concept of individualized molecular networks, called 'reticulocytes', which can be personalized further with unique exposomes, environmental exposures, and lifestyle factors, providing a more holistic understanding of disease mechanisms [29].

Computational techniques for neurodegenerative diseases

Neurodegenerative illnesses advance slowly and are exceedingly difficult to detect, posing a serious risk. In general, conventional techniques are insufficient in terms of quick diagnosis and effective monitoring. However, Advanced machine learning techniques have provided new insights into the causes of various disorders and suggest potential treatment options. One of these methods is the effective estimation of illness risk using Polygenic score-based genetic data aggregation. The use of polygenic scores and epistatic interactions for risk assessment is one of the primary objectives of precision medicine in neurodegenerative illnesses. Due to the identification of several susceptibility variations that affect the development and progression of diseases such as Parkinson's disease (PD), next-generation sequencing (NGS) technologies have provided an improved understanding of these intricate pathologies. When combined, these variants provide important insights into the genetic components of various disorders, despite the small odds ratios of the individual variants. Patients can be categorized according to their age of onset and unique endophenotype, and polygenic scores which combine the effects of many genetic variants-can be used to predict the likelihood that a disease would develop. For instance, compared to people with a later onset, those with a polygenic score higher than 1.5 are more likely to experience early-onset Parkinson's disease (PD) [30].

Clinical practices related to Alzheimer's disease (AD) and related disorders are being transformed by big data and computational resources, according to studies. Recent research highlights the process of combining many data sources, such as biomarkers, genomic profiles, cognitive tests, and high-resolution medical imaging, to create comprehensive, unique health profiles. Patterns and predictors of illness progression can be found through the use of machine learning algorithms, especially those built for handling large-scale and high-dimensional data. In order to identify modifiable risk variables and multi-determinant causal linkages, studies have emphasized the use of both supervised and unsupervised learning approaches in the analysis of these large datasets. An example of its use in the analysis of brain imaging data is the use of convolutional neural networks (CNNs), which have shown potential in precisely recognising early signs of Alzheimer's disease (AD) [31]. Using genetic data, a convolutional neural network (CNN) called ALS-Net has been developed to predict the onset of amyotrophic lateral sclerosis (ALS). Motor neurons are impacted as the neurodegenerative disease progresses, weakening muscles until they eventually collapse and result in paralysis. Their complicated genetic makeup is caused by a variety of causes, one of which is the presence of very significant genes and polymorphisms that control gene expression within promoter Because promoter regions are known to be sensitive to mutations that cause disease, ALS-Net's structure makes use of the structure of genomic data. In order to improve prediction accuracy, a two-level strategy was utilized. The model can capture intricate interactions between genetic variations that may be involved in ALS because of its structure. Research shows that when it comes to genotype-based ALS prediction, ALS-Net performs better than conventional classification methods like logistic regression and support vector machines. Massive amounts of genetic data can be processed and analyzed using ALS-Net, showcasing its potential. ALS-Net's capacity to analyze vast volumes of genomic data, with a focus on discovering genetic markers and comprehending the genetic origins of disorders like ALS, indicates the potential of machine learning approaches in precision medicine. There are also some difficulties associated with this implementation. These include the requirement for substantial computer resources, the intricacy of preparing genomic data, and the need for big, excellent datasets to train the models. Furthermore, a major obstacle still facing deep learning models is their interpretability since physicians need easy to comprehend predictions in order to make decisions. However, there are also a number of strategies to deal with these issues. Resource needs can be reduced by utilizing cloud-based platforms and improving computational infrastructure. Creating standardized processes to integrate multi-omics data and pre-process genetic data can enhance the capacity for generalization and performance of the model. Moreover, the integration of accessible artificial intelligence methodologies can aid in clarifying the decision-making procedure of deep learning models, consequently enhancing their adoption and practicality in medical environments. It highlights the wider use of deep learning in precision medicine and demonstrates the potential of convolutional neural networks in predicting ALS from genetic data. It also provides the foundation for the integration of sophisticated computational tools into clinical practice, which will ultimately improve patient care and disease prediction by addressing the issues of data complexity and model interpretability [32]. In addition to emphasizing the integration of different data types, the value of collaborative networks, and the specific applications in risk assessment and personalized treatment, this explanation offers a thorough overview of how computational techniques are being applied to neurodegenerative diseases [30].

AI for targeted drug delivery in cancer therapy

AI is increasingly paving the way for early detection of cancer using developing minimally invasive procedures such as liquid biopsies for circulating tumor DNA (ctDNA) or cfDNA. Liquid biopsies, collected through minimally invasive techniques like blood tests, can help discover cancer early, evaluate relapse risk, and guide treatment options [28]. The use of AI in analyzing highly complex datasets comprising genetic, phenotypic, and pharmacological interaction data is fundamental to the personalization of drug delivery. By precisely delivering medications to particular tissues or cellular membranes, this customized technique improves treatment success while lowering unwanted effects. Pharmacokinetics evaluation is one prominent area where AI is being used. The models use a variety of inputs to estimate the optimal therapeutic doses and delivery timings for regulated pharmaceutical release with maximum efficacy and minimal toxicity. AI is also utilized to identify and profile cancer biomarkers, which is an important procedure in prognosis, risk stratification, and therapy efficacy. In designing and functionality, the drug-loaded nanoparticles use AI to ensure exact targeting to tumor sites, reducing harm to healthy tissues and enhancing treatment effectiveness [33].

Three significant progresses have been made in the use of artificial intelligence in cancer therapy drug delivery, to improve therapeutic efficacy and optimize delivery procedures. Machine learning techniques improve the design, characterization, and manufacturing of drug delivery nanosystems by analyzing large genetic and biological datasets. This enables the rapid discovery of new drugs and accurate prediction of small molecule behavior, making AI-integrated drug delivery a key element in advancing cancer therapy. AI algorithms are capable of forecasting which treatment combinations have the best efficacy and are extremely useful in targeting tumors more precisely. By accurately localizing medications at the tumor spot, this predictive capability maximizes therapeutic advantages while minimizing negative effects. Personalized therapy approaches are made possible by the integration of AI to manage massive and complicated datasets, such as drug characteristics, patient genetic profiles, and other biological parameters. AI techniques are also used in developing an understanding and prediction of how patients would react to certain combinations of treatments, resulting in the development of more specialized and successful treatments. Pharmacokinetics and pharmacodynamics are evaluated by AI to guarantee the effectiveness and reliability of medication delivery systems. These developments demonstrate how AI may fundamentally alter the way targeted drugs are given in cancer treatment, resulting in more accurate and customized regimens [34].

Targeted cancer treatment may advance significantly with the use of AI when combined with drug delivery methods through nanoparticles. AI can optimize the functionality and design of nanoparticles (NPs), increasing their precision and efficacy in drug administration. The use of gold and mesoporous silica nanoparticles (NPs) incorporated with enzymes as drug delivery nanomotors has been recognised in recent study. Improved real-time tracking and active swarming dynamics are made possible by these radiolabelled nanomotors for in vivo imaging. This facilitates theranostic applications—a technique in nuclear medicine and personalized medicine where one radioactive drug is used to identify (diagnose) and another to treat cancerous tumors. Predicting the presence of cancer cells and optimizing nanorobot performance for targeted medication delivery are critical tasks for AI technologies, especially artificial neural networks (ANNs). After a tumor diagnosis, fuzzy logic models improve medicine dosage prediction for intracellular delivery even further. Despite its promising improvements, nanomedicine still faces several challenges, including the influence of enhanced permeability and retention (EPR), biocompatibility, medication concentration management, and potential toxicity. Nanorobots have many challenges in their therapeutic application, including fabrication-related problems, noise, and unidentified properties that alter drug dosage distribution. Artificial intelligence integration may minimize these challenges and enhance medicine formulation and dosage distribution by enhancing complex data processing and analysis. Moreover, the speed and accuracy of AI for genetic programming and pattern recognition play a considerable role in cancer genomics with optimizing efficacy of therapy and early identification of markers. Efficacy and reliability of the drug delivery system can be enhanced by expediting the study of explanatory- response variables with sophisticated AI techniques such as response surface procedure and supervised associating networks [35].

D. Data Collection and Preprocessing

Data used in building drug toxicity prediction models

Information for toxicity prediction is primarily from the drug compound’s chemical structure. These chemical structures need to be represented by numbers and characters so that they become computer-readable and interpretable ways, so-called chemical descriptors, to be effectively processed by computers [25]. The descriptor types depend on simple features, like atomic counts or molecular weights to structural features [37]. Different combinations of chemical descriptors and machine learning models might lead to different performance. Many studies have collected high-quality data from various databases on drug toxicity prediction publicly accessible, which is the critical first step for building ML models (Table 1).

Molecular descriptors, the most traditional molecular representations, have been widely utilized in toxicological prediction and other Quantitative Structure-Activity Relationship (QSAR) modeling studies combined with a variety of machine learning models. Molecular fingerprints is another broadly used molecular representation. Structural fingerprints, the most typical type of molecular fingerprints, encodes the molecular structure information into binary strings (strings of 0 and 1) [38]. The most seen molecular fingerprints include PunChem fingerprint, MACCS fingerprint , Klekota-Roth fingerprint [39], Estate fingerprint [40], and etc. Extended Connectivity Fingerprint (ECFP) is a new type of molecular representation [41], which was particularly modeled for structure-activity relationship modeling, and has a number of valuable, not pre-defined qualities. In addition, any number of different molecular features can be represented and interpreted easily with ECFP. Traditionally, quantitative structure-activity relationship (QSAR) modeling studies were used for computational drug toxicity prediction [42]. QSAR has been used to study the quantitative relationship between molecular structure and biological activity [43]. On the other hand, the prediction model which was obtained by combining

Database Name	Description	Reference
TOXNET	-Computerized system of files relevant to toxicology and related areas. -The world's largest collection of toxicology databases -Available files: HSDB (Hazardous Substances Data Bank), TRI (Toxic Chemical Release Inventory), IRIS (Integrated Risk Information System), etc	Wu Y & Wang G 2018 [25]
ToxCast	High-throughput toxicity data on thousands of chemicals - Based on HTS assays, cell-based phenotypic assays, and genomic and metabolomic analyses of cells. - Providing a standard for consistent and reproducible data processing for diverse, targeted bioactivity assay data	Dix et al. 2007 [46]
admetSAR	Accessible high quality datasets about absorption, distribution, metabolism, excretion, and toxicity (ADMET) - Containing over 210,000 ADMET annotated data points - open source, text and structure searchable, continually updated	Cheng 2012 [47]
PubChem	The world's largest collection of freely accessible chemical information - Including physical properties, biological activities, safety and toxicity information, etc - Name, molecular formula, chemical structure, and other identifiers searchable - Used in many ML	Kim 2021 [48]
ISSTOX	Chemical toxicity databases Containing experimental results relative to various types of	Benigni et al. 2013

	chemical toxicity - Characterized by the use of contemporary information technologies and by organization of high-quality biological data	[49]
ChEMBL	An open large-scale bioactivity database - Containing binding, functional and ADMET information for a huge number of drug-like bioactive compounds - 5.4 million bioactivity measurements for over 1 million compounds and 5,200 protein targets	Gaulton et al. 2012 [50]
BindingDB	One of the most extensive public-accessible databases of protein-ligand binding affinities - Currently holding ~20,000 measurements for ~11 000 different small molecule ligands and 110 different drug−targets	Liu et al. 2007 [51]
T3DB	The Toxin-Toxin-Target database containing information about the toxic exposome - Currently corrected and upgraded to include more compounds, targets, gene expression datasets along with extensive toxic compound concentration or exposure data.	Wishart et al. 2015 [52]
DrugBank	A bioinformatics and cheminformatics resource combining sophisticated drug data with comprehensive target information - Characterized by the richness, uniqueness and quality of its data - Including illustrated drug-action pathways, drug transporter data, drug metabolite data, pharmacogenomic data, adverse drug response data, ADMET data, pharmacokinetic data, extensive computed property data and chemical classification data	Knox et al. 2011 [53]
ECoTOX	A reliable source of curated, currently updated, ecological toxicity data for chemical assessments and research - Providing single chemical ecotoxicity data for over 12,000 chemicals	Olker et al. 2022 [54]
SuperToxic	A comprehensive database of toxic compounds - Collection of toxic compounds compiling about 60,000 compounds and their structures - Classified depending on their toxicity, based on over 2 million measurements	Schmidt et al. 2009 [55]

Table 1. Publicly accessible databases for drug toxicity prediction

machine learning and the molecular descriptors is similar to QSAR toxicity but including environmental behavior of chemicals [44,45].

Besides chemical structures, machines should fully exploit gene expression data and transcriptome expression data for feature selection and classification in drug trials because not only drugs are designed for humans, but also those data reflect the state shifts of a cell, either in vivo or in vitro [56]. Drug toxicity can be predicted by the analysis of changes in the transcriptome [57]. Moreover, analyzing gene interaction networks allows one to predict the delayed chemical toxicity as well as to be provided with richer biological information. Although the number of affected genes is small at the induction, much greater gene expression changes will take place 24 hours after induction [58]. In other words, as gene expression doesn't occur immediately, toxicity onset is often delayed and it is difficult to identify immediately after the induction. Therefore, the degree of toxicity would positively correlate with the degree of connectivity of the genetic network [59,60]. Incorporating genetic information has a few more advantages in terms of the accuracy of toxicity prediction and QSAR construction [61]. When it comes to biological information, it is not only at the molecular level(involving only a single pair of drug-protein interaction), but at the systems level with a drug targeting the entire gene interaction network. Furthermore, one is able to distinguish between toxic and non-toxic and perform classified toxicity prediction.

However, there are still a few limitations in the use of computational techniques in toxicology despite their significance. One of the main challenges is the lack of ability to obtain a mechanistic explanation or understanding of the detected toxic responses. The ML algorithms are generally considered as back boxes which are capable of managing complicated problems efficiently but often lack an explanation of the prediction [62]. In addition to difficulties in interpretation, it is difficult to precisely estimate the prediction performance of the model without using cross-validation [63,64,65]. Unless the model repeats enough times for building a large number of models, the accuracy of the estimated model may be biased.

Many studies have applied machine learning methods in drug toxicity prediction, such as diverse toxicity endpoints, such as an carcinogenicity, mutagenicity, hepatotoxicity, acute oral toxicity, and human ether-a-go-go-related gene (hEGR) inhibition [38], as mentioned above, with varied types of datasets like molecular descriptors and fingerprints. Following are the recent drug prediction models integrated several ML methods with molecular descriptors and data including a number of chemical compounds reported in the literature.

Carcinogenicity: Helma et al. [66] developed a rodent carcinogenicity prediction model (named lazer) in order to predict carcinogenicity of diverse chemicals using a modified k-nearest-neighbor (knn) algorithm and the Lois Gold Carcinogenic Potency Database (CPDB), which contains findings of 2-year rodent carcinogenicity study for 1481 chemicals with diverse chemical structures. Making use of the same database and the twenty-seven two-dimensional MDL descriptors as molecular representation, Fjdorova et al. [67] developed counter propagation artificial neural network (CP ANN) models, and two public carcinogenicity prediction models using 8 MDL descriptors and 12 Dragon descriptors based on these algorithms.

Mutagenicity: The Ames mutagenicity benchmark dataset developed by Hansen et al. [68], which contains 2503 positive compounds and 3009 negative compounds(6512 in total), is the most commonly used training dataset for the development of predictive compound mutagenicity models. Xu et al. [69]

established a series of mutagenicity prediction models using five machine learning algorithms (SVM, knn, naive Bayes, ANN and decision trees) and a data set with 7,617 compounds.

Hepatotoxicity: A hepatotoxicity prediction model developed by Ekins et al. [70] is based on the Bayesian approach utilizing ECFP molecular fingerprinting. The model was trained on a training set of 295 compounds and tested on a test set of 237 compounds. Zhang et al. [71] modeled a series of hepatotoxicity prediction models applying SVM in calculated three types of molecular fingerprints for 1229 compounds and Naive Bayes algorithm for the specificity improvement.

Acute Oral Toxicity: Li et al. [72] developed several multiclassification models for acute oral toxicity prediction using five machine learning methods based on a data set including 12204 diverse compounds with LD50 values.

hERG inhibition: Zhang et al. [73] collected a hERG blockage database containing 1570 compounds and established several classification models for hERG inhibition prediction with five machine learning and molecular descriptors combining fingerprints.

Traditional and AI-based approaches for data collection in Cardiovascular diseases

AI-driven approaches have revolutionized the approaches to cardiovascular diseases (CVDs) including coronary artery disease (CAD), heart failure, arrhythmias, and congenital heart defects [28]. Due to the complex and heterogeneous nature of CVD, the traditional paradigm, which mostly a fixed set of observation variables initially and a considerable length of period till the outcomes collection are required, used for building risk models from a population-based study faces a severe challenge to the development of accuracy [74,75]. Research in CVD has mostly persisted for over a century and progressed with reductionist methods that deconstruct the problem into its constituents, investigate the parts independently to get insights on associations and causal relationships. However, this reductionist approach remains an inability to uncover a complex genotype-phenotype relationship [76]. Reductionism professes that a pathogenetic variant functions as the primary determinant of a disease trait or endophenotype, which is a critical step when it comes to development of a clinical disorder. Rather, the expression of overt cardiovascular end-pathophenotypes more likely reflects the combined and interacting effects of perturbations in a number of potentially phenotypically related genes that are modified by an individual's exposome, the cumulative environmental exposures that affect health [77]. Due to this feature, GWAS used to dissect complex disorders like CVD has several limitations. GWAS are limited by the fact that the genetic heterogeneity of many variants is common and also sensitivity is still restricted by the depth and coverage of the sequencing platform. In addition, GWAS are only capable of providing an association between geon regions that the pathogenic gene might reside in and the disease phenotype. Insufficiency of population also restricts the statistical power needed to discover even simple gene-gene interaction, even for highly prevalent CVDs (Figure 3) [78].

On the other hand, AI-based technologies aim to build algorithms for better accuracy and prediction of heart-related disease by combining these AI and ML technologies with a large group of health records with extensive variety, high volume, and utmost acquisition speed (e.g human gut microbiome sequencing, multi-dimensional data, social media, and data from standardized electronic health records (EHRs), precision medicine platforms and data from wearable technology [79,80,81]. Current understanding of the heterogeneity in cardiovascular disease requires compilation of a variety of big data sources. Data collected from these domains are compliant to novel network medicine analytics to generate individual patient networks based on population-level data as well as reticulotypes (individuals' unique genomic and molecular makeup). Network medicine facilitates precision endophenotyping and phenotyping for patients with similar clinical symptoms. Utilizing big data, patient-specific integrated networks can be modeled, and the consequences of perturbations occurring due to reticulotypes can be scouted. The reticulotype also controls endophenotype and defines a patient-specific phenotype that might not have been apparent before [27].

Cardiovascular diseases are compilations of complicated clinical phenotypes including a number of different endophenotypes such as inflammation, calcification, thrombosis, fibrosis. That is, these complex endophenotypes cannot be explained by just a single genetic variant. The classical model based on reductionism purports a disease trait or endophenotype can be determined by a pathogenic gene variant present in GWAS. (B) Contemporary viewpoint of CVDs using big data sources enables not only to create patient-specific integrated networks or reticulotypes but also defines a patient-specific pathophenotypes and treatments. Moreover, there are some reliable websites mostly used for collecting data of patients, such as UCI and Kaggle. UCI Machine Learning Repository and Cleveland Heart

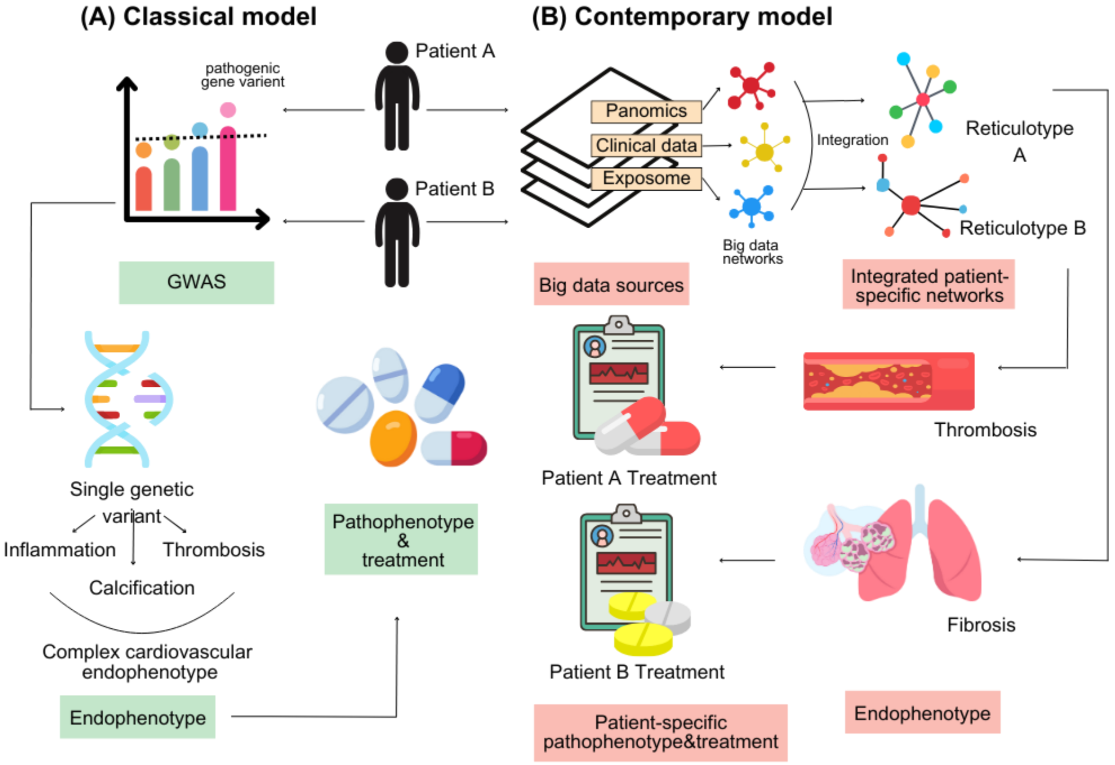

Figure 3. Classical and contemporary models processing data towards the treatments

Disease dataset, which is acquired from Cleveland Clinic Foundation, Cleveland, OH and provided by Robert Detrano, M.D., Ph.D. of the V.A. Medical Center, Long Beach, CA, contains 76 features and 303 instances The UCI repository dataset, especially the Hungarian and the Cleveland data sets, is promising in terms of the number of data instances and quality of data [82]. As an example, David and Belcy conducted a comparison study on heart disease with RF, DT, and NB to build a prediction method for examination and prediction of the potential of heart disease. The statlog dataset from the UCI repository, which included 270 cases and 13 attributes, was used for the model training and testing. Nandhini et al. developed a model for real-time prediction of cardiovascular disease [83]. They utilized the dataset from UCI repository of common Cleveland HD datasets containing 303 cases and 76 gestures to train and test ML techniques such as NB, SVM, RF, LR, KNN and DT.

There are some challenges and limitations that remain to be addressed in the implementation of a big data approach in cardiovascular medicine. First of all, integrating a less controlled database into clinical trials by using traditional methods is difficult due to the ideal conditions(e.g. select patients and experienced physicians) [84]. Second, heterogeneity and disparities among datasets pose utilization challenges, while latent variables like hidden medical and lifestyle factors may have been overlooked in previous studies [85,86]. Lifestyle variables, often novel, are challenging to integrate due to data privacy issues and lack of public available application programming (API) for consumer devices to interact with electronic health records (EHRs) [87]. Wearable technology advancements could track real-time lifestyle factors, which could exclude the possibility of recall or social desirability biases [88], but issues with FDA approval, validation, and long-term behavioral change motivation remain [89,90]. Third, data quality, inconsistency, instability, and validation issues are barriers necessitating critical data imputation for big data analysis [91]. Fourth, synchronizing existing data can be challenging because of heterogeneity of multiple databases (i.e., different diagnostic criteria, different laboratories) [92,93]. Fifth, data privacy issues may remain unsolved with de-identification because there are various ways that re-identification can be performed [94]. Sixth, genomic data, or GWAS, may still need human intervention [95] as there's insufficient evidence that DNA testing has a significant impact on motivating behavior change [96]. Last challenge, which is so important that all big-data analyses must take into consideration, is the ascertainment of causality from observational and retrospective studies. This is because most AI/ML methods are not clearly designed for modeling causality. For example, although gray hair, wrinkles, baldness can be presented by both aging and CVD and are highly correlated, the therapies for aging (e.g. hair dyes or wrinkle cream) and the treatments for CVD are completely different. That is, if we only pursue this association and design therapies for aging, we would totally fail to prevent CVDs [97,98,99].

ML is currently applied in a variety of areas in cardiology, including diagnostic imaging, electrocardiography (ECG), patient cardiovascular risk assessment and prognosis prediction. For instance, Madani et al. [100] trained a convolutional neural network (CNN) to recognize 15 standard echocardiographic views, using a training and validation set of over 200,000 images and a test set of 20,000. It outperformed board-certified echocardiographers with an overall accuracy of 91.7%. The study conducted by Galloway, et al. [101] used ML for the screening of hyperkalemia in patients with chronic kidney disease from ECG derived from three Mayo Clinic centers in Minnesota, Florida, and Arizona using an analysis of a database of 449,380 patients of different hospitals and obtained high sensitivity.

Data Collection for application of ML in neurodegenerative disease

Neurological disorders, such as stroke, Alzheimer's disease, Parkinson's disease, epilepsy, and Amyotrophic lateral sclerosis (ALS), raise difficulties, on account of their complex etiology, heterogeneous presentation, and variable response to medication. The employment of AI and ML technologies into precision medicine has shown the potential in analysis of various types of biomedical big data. A large volume of data including electronic health records (EHRs), genomic profiles, imaging data, wearable sensor data and multi-omics [27] has been rapidly produced, as production of a wide variety of biomedical data getting simpler and faster [102], to tackle its heterogeneity, driven first by genomic studies, and to date transcriptomic and epigenomic studies [27]. Omics refers to the comprehensive characterization and quantification of molecules such as genes and proteins clustered depending on their structural or functional similarities, and include genomics, proteomics, metabolomics, etc. Multi-omics data integration combines information from different layers of omics data to obtain understandings regarding how different biological systems interact at a molecular level [103,104], which is a crucial step to recognize inherent pathogenic pathways in different disease phenotypes. It is pertinent in neurodegenerative diseases (NDs) such as AD and PD, which involve a multifactorial etiology with heterogeneous clinical pictures and mixed pathologies [105]. EHRs are composed of patients' demographics along with clinical measurements, interventions, clinical laboratory tests, and medical data. EHRs can be divided into structured, which are represented by diagnostic codes and laboratory test outcomes, and unstructured, which are represented by patients' status annotated by physicians. Classical statistical methods are considered not suitable for analysis of these kinds of data and more advanced technologies are required. Deep learning models particularly can be exploited to find patterns in a patient's broad scope view and search explanations of differences between groups as hypothesis-free methods, compared to classical experiments with hypothesis [102].

Besides the traditional and prevalent issues such as overfitting and interpretability, one of the major issues in big data analysis is data isolation, posing considerable challenges for researchers to access, integrate, and construct noisy, complex, and high-dimensional data. This is because healthcare data are typically dispersed across various medical systems and many of these are not interconnected, resulting in isolated data and the rise in the expenses of institutions. It impedes the healthcare entities from leveraging the latest Information Technologies (IT) innovations including data processing and cloud computing, which might be helpful for improved care and reduced costs [106]. Also, there's a concern with regard to the reproducibility of other studies and the implementation of others' AI models. This occurs owing to limited open-source implementations provided by authors and the difficulty of recreating a network in a different software [107]. Lastly, data sparseness in computational diagnosis and treatment remains as an unsolved challenge. Sparse features, which include most zero values and relatively few nonzero values, in data are inefficient as they take up computing memory and reduce generalization ability [108].

A deep learning based model using genome data for ALS that was designed by B. Yin et al showed promise of DNNs. Most heritability of ALS hasn't been explained so far while several major genetic risk factors have been identified. This is because it cannot be detected using the currently available genotype-phenotype association approaches [109]. As a result of the recent advances, deep neural networks (DNNs) have proven to be powerful classifiers in several applications including bioinformatics [65] due to its ability to handle a significantly larger number of input variables than most other ML methods [110], a prerequisite for the analysis of genome data.

However, several challenges need to be overcome for DNNs to map genetic variants to disease (ALS) status. First, the size of genome data (which amount to usually millions) is numerous for these models to deal with easily. Second, interpreting the reason behind a DNN's classification of a sample as a case or control which is a primary drawback in genotype-phenotype association. Third, the employment of convolutional filters that utilizes the position invariance of local structures in images, which makes it appropriate for image classification. But when it comes to genome data, it doesn't have local structure, and applying convolution is less straightforward [32].

In order to predict the occurrence of ALS from individual genotype, B. Yin et al developed a deep learning based approach for the classification of ALS patients versus healthy controls from the Dutch cohort of the Project MinE dataset, a global effort to collect whole-genome data for the identification of ALS-causing variants [111].They validated their approach using GWAS, the cutting-edge in analyzing genotype-phenotype data [112]. They presented Promoter-CNN (Promoter region-specific neural net) + ALS-net (the network that classifies samples based on a combination of promoter regions) and observed a few improvements over logistic regression, which is typically used to predict continuous outcome scores [112]. Using GWAS, they established a two-step approach and applied it to neural network technologies for identification of associations between genotypes and the occurrence of ALS. The first step comprises the classifying process for each promoter region, in which they assume that the majority of variants related to disease phenotypes inhabit ahead genes. They only selected eight best performing promoter regions based on individuals' genomic data from a single promoter region. Afterwards, in the second step, an overall classifier trained for final classification is combined with the genomic data from selected promoter regions. They improved the drawback of earlier studies that epistasis is overlooked and variants with a small effect on their own will not be incorporated in further analysis, and illustrated the promoter regions on the different chromosomes interact in a non-additive manner with their model. Additionally, ALS-Net outperforms all other methods in terms of recall and classification accuracy, and provides a better trade-off between precision and recall. The architecture of ALS-Net, which was originally optimized for chromosome 7, also performed excellently when applying it to the classification of samples from genotype data from other chromosomes with no need for adjustment. This shows the promising impact of this architecture to be applicable more universally [32].

Generating Datasets for Cancer Predictive Models

AI and machine learning have made significant contributions to the critical area of oncology, and the main determinants of tumor aggressiveness, clinical decision-making and outcomes such as timing of cancer detection, accuracy of cancer diagnosis and staging, benefit from these approaches. In past decades, efforts for generation of large cancer datasets have been made by several initiatives worldwide. These datasets for building predictive models, which are used for informing both research and clinical decisions, are obtained from profiling tumor samples using diverse high throughput platforms and technologies. The Cancer Genome Atlas (TCGA) is the most comprehensive compliance of tumor profiles which are publicly available and includes numerous data types encompassing genomics, epigenomics, proteomics, histopathology and radiology images [113]. The Pan-Cancer Analysis of Whole Genomes (PCAWG), METABRIC, and GENIE have also curated a large number of cancer genomic profiles which the public are able to access. Besides these datasets, there are several other technologies that have resulted in the production of a wide ranging array of datasets, such as DNA methylation profiles, large scale proteomics studies, perturbation studies containing cell viability or cytotoxicity assays using small molecules, RNA interference (RNAi) or CRISPR screens, protein-protein interaction networks.

In machine learning workflows, the initial and crucial phases involve data collection and cleaning, as the quality of a model is directly dependent on the quality of the training data. For verification of high quality of the collected data, it needs to be corrected for potential noise in both non-image (e.g. erroneous data entries, missing values) and image (e.g. high intensity pixels) types of data. Second, the data needs to be reviewed for possible biases that might cause underfitting the model, or high variance that might cause overfitting the model. Third, the model needs to be measured using the Area Under the Receiver Operator Curve (AUC) or the Area Under the Precision-Recall Curve (AUPRC). They are used for quantification of the tradeoff between sensitivity and specificity, which are required to be considered as a good classifier. Fourth, it is crucial to confirm the model on external independent datasets to ensure that the model is stable and generalizes well. The model is required to keep being tested from time-to-time as newer updated data sets become available to prevent the performance of it from degrading due to concept drift. In other words, it is required to show stability when the relationship between the input and output variables changes over time unpredictably [113].

Deep Learning-based models have been reported to have potential not only in the accurate diagnosis and identification of cancer subtypes directly from histopathological images, other medical images (those acquired through non-invasive techniques such as Computed Tomography (CT) scans, Magnetic Resonance Imaging (MRI) and mammograms) and even images of suspicious lesions , but in cancer staging and grading. As an example, Coudray et al. [114] developed and applied DeepPATH, Inception-v3 architecture-based models to classify H&E-stained whole slide images (WSI) for the TCGA lung cancer cohort into three classes (normal, lung adenocarcinoma and lung squamous cell carcinoma) with high AUC. This research demonstrated DNNs are powerful enough to be used for distinguishing between closely related cancer subtypes and detecting benign vs. malignant tissue (i.e. non-cancerous vs. cancerous tissue), which is regarded as a challenging task. Cancer staging and grading is the process that deminines how aggressive and advanced the observed cancer is, and another crucial component throughout the diagnostic process. DNNs have shown promising results in predicting Gleason scores, a combination of two scores measuring how prevalent tumor cells are in two distinct locations on a slide, from histopathology images of prostate tumors, radiology images, and increasingly non-imaging data (such as genomic profiles) obtained by next generation sequencing (NGS) [35]. However, DNN has a limitation in terms of multi-class classification using mutation data. For example, Sun et al. built and applied DNN to genomic point mutations to classify tissues into either of the 12 TCGA cancer types or healthy tissues obtained from the 1000 Genomic Projects [115]. The classifier was trained on the most frequent cancer specific point mutations obtained from whole exome sequencing profiles, and successfully distinguished between healthy and tumor tissue with high accuracy (AUC=0.94), but did not perform as well in a multi-class classification task to distinguish all of 12-cancer types at the same time.

The integration of AI with clinical interventions necessitates the fulfillment of several specific conditions. First of all, data should adequately encompass the entire human population. In cancer, it has been reported that race-specific variations influence the occurrence and frequency of genomic aberrations. Existing datasets commonly used for AI model training and testing still contain inherent bias towards particular race and ethnicity. As an example, in TCGA, the largest repository of diverse cancer datasets, white individuals with European ancestry take the most part of it [116]. Aside from data biases, even though data can be acquired from various platforms, external institutions are limited to access the data for independent use, especially for private or controlled access data sets. Moreover, as to code sharing, the models of published studies should be reproducible by others independently to verify they are translatable and clinically relevant. This can be done by sharing well annotated code for the model with clear descriptions of the optimized hyperparameters and hardware specification, which is not adopted generally. Also, while AI cancer models currently used have a strong impact on image and omics data, the huge part of electronic health records (EHRs), which cover one of the richest data of patient health and clinical history, hasn't been fully utilized. This is because records not being organized with high levels of noise, sparseness and inconsistencies, that is, committed curation and data cleaning are required [35].

A. Case Studies

Case Study 1: Computational precision medicine for Alzheimer's and other neurodegenerative illnesses

A study by Zang et al, 2022 utilized a high-throughput clinical trial simulation technology and real-world data to simulate 430,000 Alzheimer's disease medication repurposing trials, including **propensity score-based** causal inference. Repurposing trials examine current medications for new therapeutic objectives, with the goal of discovering new applications or advantages that go beyond its original usage. Propensity score-based causal inference is a statistical approach for estimating causal effects in observational research by balancing covariates across treatment groups, hence eliminating bias and confounding effects. It identified eight medications that might assist Alzheimer's patients and emphasized the need of a model selection technique in enhancing confounding balance in large-scale studies. The study revealed eight medications with various initial indications that may assist Alzheimer's disease patients. The suggested model selection technique considerably enhanced confounding balance performance in simulated trials, highlighting the significance of model selection in large-scale trial balancing. A **regularized logistic regression-based propensity score model** outperformed deep learning-based models in trial balancing, due to its effective use of propensity score-based causal inference. This statistical method assists in balancing variables between treatment groups, reducing bias and confounding factors and improving model accuracy. This case study demonstrates the practical application of computational approaches in precision medicine for neurodegenerative illnesses, highlighting the importance of model selection in generating balanced trial results [117].

Case Study 2: Integrating Computational Innovation for Parkinson's Disease

Unlike previous techniques, which frequently depend on single biomarkers or limited data kinds, Makarious et al.'s 2022 study combined many data modalities. This comprises genetic data, clinical records, imaging data, and other omics data (such as proteomics or metabolomics) from PD patients without any other additional neurological disease or retraction toward PD diagnosis during follow-up; this set of data was collected in collaboration with the AMP PD and GP2 initiatives of PD patients. Data from the PPMI and PDBP cohorts were included, considering only those characteristics available in at least 80% of the training and validation cohorts. Analyses were restricted to unrelated individuals of European ancestry. Standard Illumina technologies were employed to generate DNA and RNA sequencing data, while simultaneously applying rigorous quality controls. Regression models with various normalization techniques have been used to select these features to fit data. Respectively, multiple ML algorithms were used to stick to a model supplied by performance measures on test data for the identifications of performances using AUC, balanced accuracy, sensitivity, and specificity. The feature selection was based on the extraTrees algorithm to avoid overfitting and redundancy. Cross-validation techniques tuned the best-performing algorithm, which was then further validated with the PDBP dataset. Further, a post hoc optimization was implemented to adjust the probability threshold to improve model accuracy in unbalanced cohorts. As the last phase of the process, Shapley values were used to interpret the importance of each feature in the model and make it interactive to find out the contribution of features at the individual level in classification. Makarious et al.'s 2022 study represents a major advancement in neurodegenerative illness research, with an emphasis on Parkinson's disease (PD). The study highlights the potential for computational approaches to revolutionize early diagnosis and intervention efforts by combining varied data sources with powerful machine learning techniques. The study's goal in merging these varied sources was to capture a more comprehensive perspective of the condition, maybe uncovering subtle patterns or interactions that would not be apparent from a single source. The study created effective computational models for the early detection of Parkinson's disease. These models most likely used machine learning methods that can handle multidimensional data and discover complicated patterns across several data types at the same time. These models not only enhance diagnostic accuracy but also allow for earlier identification, which is critical for beginning early therapies. Beyond diagnosis, the study investigated gene networks and drug-gene interactions in Parkinson's disease. This method can provide new biomarkers, therapeutic targets, and personalized therapy alternatives. Understanding how genes interact with one another and with drugs might result in more effective therapy options customized to patients. The research is significant for its emphasis on cost-effectiveness and transparency. The authors reduced data collecting and analysis expenses by relying on publicly available data and automated machine learning techniques. Furthermore, they openly released their code and findings, encouraging reproducibility and cooperation within the scientific community. This study's findings have implications that go beyond Parkinson's illness. Similar computational techniques might be used to treat other neurodegenerative disorders, such as Alzheimer's, Huntington's, and amyotrophic lateral sclerosis. By utilizing large-scale data integration and sophisticated analytics, researchers can obtain deeper insights into disease causes and accelerate the development of diagnostic and therapeutic advances [118].

Case Study 3: Advancing Predictive Models for Chronic Heart Failure Readmission Risk

Liu et al. 2022, examined prediction models for readmission risk in Chronic heart failure (CHF) patients. This case study demonstrated how predictive models based on big data might estimate the probability of readmission in patients with Chronic Heart Failure. The study includes nine CHF readmission prediction models, each with a different sample size and modeling approach, including logistic regression, Cox proportional hazards, competitive risk, ensemble learning, and Bayesian models. The initial Area under the Receiver Operating Characteristic Curve (AUROC) evaluations (0.70 to 0.73) in the study indicate that the predictive models were good at distinguishing between CHF patients who would be readmitted and those who would not, implying that they were better than random guessing but still had potential for improvement. The AUROC is a metric used to evaluate the performance of predictive models. Typical values range from 0.5 (random guessing) to 1 (perfect prediction). In the study, initial AUROC evaluations of 0.70 to 0.73 indicated that the predictive models were better than random guessing but had room for improvement. After verification, the AUROC scores increased to 0.80, demonstrating strong predictive accuracy. An AUROC score of 0.80 implies that the revised models were effective in predicting readmissions, providing medical professionals with better decision-making choices. After verification, the AUROC scores increased from 0.68 to 0.80. These models were examined using meta-analysis approaches to determine the predictive value of variables. The top end of this range (0.80) implies strong predictive accuracy, demonstrating that the revised models were more effective in predicting readmissions, providing medical professionals with greater decision-making choices. To ensure the accuracy of the results, statistical approaches such as odds ratio (OR) or risk ratio (HR) with 95% confidence intervals were used, as well as sensitivity analysis. The objective of this case study in evaluating these models was to give insights into improving the prediction of readmission risk for CHF patients utilizing data-driven precision medicine techniques [119].

Case Study 4: Enhancing Cardiovascular Disease(CDV) Prediction through Big Data Integration

The study by Guo et al. 2021, focuses on developing a risk prediction model for incident heart failure using machine learning approaches that are specifically targeted to the African American community. It used data from the Jackson Heart Study (JHS) database and used several imputation procedures to create an ideal prediction model using algorithms such as K-Nearest Neighbour, random forest interpolation, and, most significantly, XGBoost. XGBoost performed better, especially when missing rates were less than 30%, demonstrating that improved imputation approaches may dramatically improve prediction accuracy in datasets with partial information. XGBoost outperformed other machine learning algorithms, particularly when dealing with datasets containing missing values, due to its unique handling of missing data and strong imputation procedures. A comparative review of machine learning techniques such as logistic regression, support vector machines (SVM), AdaBoost, and XGBoost indicated that XGBoost beat the rest, with an outstanding Area Under the Curve (AUC) of 0.8409 for predicting heart failure in the JHS cohort. To achieve that, after imputations were performed on the JHS data, categorical variables were transformed by one-hot coding and continuous variables were normalized using Min-Max normalization. This data set was then split 70% for training and 30% for testing. A class imbalance problem was recognized, which was handled by adjusting the "class weight" parameter in models such as LASSO logistic regression, SVM and random forest. For the tree-based XGBoost algorithm, the "scale_pos_weight" parameter was adjusted to improve predictive ability, using K-fold cross-validation. Moreover, the algorithm tuning process comprises carrying out a set of parameters to reduce the learning rate and add more trees to enhance the algorithm accuracy. This demonstrates the effectiveness of gradient boosting approaches in collecting complicated data linkages required for good prediction. The study identified certain characteristics, such as variations in diabetic treatment, as important predictors of heart failure risk. These variables continuously showed substantial connections to the result across various imputation procedures, emphasizing the necessity of careful feature selection and domain understanding in developing models. The study's emphasis on the influence of imputation procedures and variable selection gives practical insights for enhancing predictive modeling in healthcare settings. While the study focuses on the African American population in the JHS cohort, its techniques and conclusions may be applicable to other demographic groups and datasets with comparable features. Future study should investigate the models' transferability across different demographics and confirm their usefulness in a variety of medical settings. Using these approaches, healthcare practitioners may be able to reduce the burden of heart failure by improving the identification of at-risk patients and implementing focused intervention measures based on unique patient needs [120].

Case Study 5: TargeTox: A Machine Learning Approach for Optimised Drug Toxicity Prediction

The study by Lysenko et al. 2018 focuses on the development and use of a machine learning method called TargeTox, and it predicts drug toxicity in precision medicine. TargeTox, the proposed machine learning algorithm, uses drug targets, off-targets, functional impact scores, and biological network data to estimate drug toxicity. The study gathers information on pharmacological targets, off-targets, and biological networks from many databases and pieces of academic literature. Functional impact scores assess the biological importance of drug-target interactions. It also uses data from biological networks to better comprehend the overall impact of medication interactions in the cellular environment. TargeTox introduces and analyses acquired data using machine learning techniques, discovering drug toxicity-related patterns. The model is trained using a set of known toxic and non-toxic drugs. Idiosyncratically toxic drugs cause unpredictable, uncommon side effects that are not dose-dependent and are connected to off-target effects. The study shows that the approach can discriminate between idiosyncratically toxic and safe medications, which helps anticipate drug toxicity. TargeTox can be used during the early stages of drug development to identify possible harmful compounds. It is very helpful to create medication combinations with lower toxicity depictions, supporting precision medicine by customizing medication treatment based on individual toxicity profiles while decreasing side effects. Additionally, it becomes a useful tool for finding potentially dangerous substances and constructing low-toxicity combinations from its use of drug targets, off-targets, and biological network data as mentioned previously. This clearly demonstrates the practical application of machine learning in drug toxicity prediction. It provides a helpful tool for anticipating drug toxicity, hence improving the safety and benefit of pharmaceutical treatment in precision medicine [121].

Case Study 6: Utilizing Machine Learning for Enhanced Drug Toxicity Prediction with eToxPred

A 2019 study by Pu et al evaluates the use of machine learning approaches in medication toxicity prediction, specifically eToxPred, in the context of precision medicine. eToxPred uses molecular fingerprints (numerical representations of molecular structures) to assess the toxicity and synthetic accessibility of small organic compounds, with up to 72% accuracy in toxicity prediction and a mean square error of 4% in synthetic accessibility estimation. eToxPred's excellent toxicity prediction accuracy (72%) demonstrates its reliability in identifying harmful substances early in the drug development process. Its capacity to assess synthetic accessibility aids in the identification of molecules that are both safe and practical to synthesize in a laboratory environment. It combines data from a variety of sources including chemical databases and biological testing and refines its prediction models using a training dataset of recognised compounds with established toxicity profiles. Other methods, such as Pred-hERG and ProTox, are also emphasized for their effectiveness in predicting cardiac toxicity and rodent oral toxicity. Pred-hERG detects cardiac toxicity by inhibiting hERG potassium ion channels, resulting in a high accurate classification rate of 0.8 and multi-class accuracy of 0.7 2. ProTox accurately predicts rodent oral toxicity and adverse medication responses, exceeding commercial software with sensitivity, specificity, and accuracy of 0.76, 0.95, and 0.75, respectively. The use of eToxPred in virtual screening methods helps to filter down the large number of prospective drug candidates to those with the best chance of success. eToxPred provides a double benefit in the drug development process by predicting toxicity as well as synthetic accessibility, guaranteeing that only safe and manufacturable compounds proceed. Despite their great accuracy, eToxPred and other comparable systems rely largely on the quality and depth of their training datasets. Incomplete or biased data can have an influence on the reliability of predictions. Since biological systems are so complicated, no prediction model can be completely flawless, emphasizing the importance of ongoing development and validation of these techniques. As AI technology advances, tools such as eToxPred will become increasingly important in the development of future treatments [122].

Case Study 7: AI-driven Innovations in Nanoparticle Drug Delivery for Cancer Therapy

Study by K. Vora et al. 2023 demonstrates the practical application of AI in precision medicine for targeted drug delivery in cancer therapy. The major goal of the study was to increase knowledge and efficacy of nanoparticle-based medication delivery systems in cancer treatment. They used artificial intelligence in combination with physiologically based pharmacokinetic (PBPK) models to simulate and predict nanoparticle pharmacokinetics and biodistribution. PBPK models are complex tools that employ mathematical descriptions of physiological processes to forecast drug movement in the body. Integrating AI into these models enables more accurate forecasts and optimisations. Large datasets were analyzed using AI algorithms, including the biological and chemical aspects of nanoparticles. These algorithms helped in predicting how various nanoparticle formulations would behave in the body, therefore determining the best qualities for successful cancer targeting. The size of nanoparticles, as well as their surface charge and surface changes, were shown to be effective in targeting tumors using nanoparticle formulations. Nanoparticles of optimized sizes, typically in the 10-100 nm range, can aggregate efficiently in tumor tissues due to the EPR effect. Surface charge is extremely significant, and a modestly positive surface charge can improve cellular absorption. Surface modification by biological ligands such as antibodies or peptides results in specific binding to tumor cell receptors, allowing for targeted administration. The combination of AI and PBPK models dramatically enhanced our knowledge of nanoparticle biodistribution, resulting in more effective cancer targeting. The study discovered that AI-driven models could predict and optimize nanoparticle distribution to tumor areas more precisely than traditional techniques. AI algorithms were important in optimizing the physicochemical characteristics of nanoparticle compositions. The study found that AI could accurately predict drug-drug interactions (DDIs), which is important for creating combination medicines in cancer therapy. This predictive capability optimizes the drug development process, enhancing safety profiles and reducing time-to-market for novel medicines. AI models anticipate medication behavior with great accuracy, allowing for more specific targeting of cancer cells. This precision improves therapeutic results because medicines may be tailored to attack cancer cells while minimizing adverse effects on healthy tissues. AI improves research and development procedures, cutting costs and time connected with medication development. It aids in experimental design, lead compound optimisation, and minimizes the need for lengthy animal testing by precisely predicting pharmacokinetics and toxicity. As AI technology advances, its applications in targeted administration of drugs are expected to grow offering up possibilities for innovation in cancer therapy [123].

Case study 8. Deep neural networks (DNNs) for prostate cancer discovery

Advances in interpretability of machine learning models made in recent years enable discovery and prediction in clinical cancer genomics, which was verified by the study conducted by Haitham et al. (2021). Haitham et al. developed a deep-learning predictive model that incorporates previous biologically established hierarchical knowledge in a neural network language for cancer state prediction in patients diagnosed with prostate cancer based on their genomic profiles. A set of 3,007 curated biological pathways were used to build a pathway-aware multi-layered hierarchical network (P-NET) and a set of 1,013 prostate cancers (333 castrate-resistant prostate cancers (CRPCs) and 680 primary cancers) were used to train and test P-NET. P-NET is a neural network algorithm that encodes various biological entities into a neural network language consisting of consecutive layers (i.e. features from patient profile, genes, pathways, biological processes and outcome) with customized connections between each of layers. The trained P-NET outperformed typical machine learning models including decision tree, random forest, and linear and radial functional support vector machine (area under the receiver operating characteristic(ROC) curve(AUC)=0.93, area under the precision-recall curve (AUPRC)=0.88, accuracy=0.83). Moreover, P-NET demonstrated its capability of generalization when classifying unseen samples by achieving 73% true-negative rate (TN) and 80% true-positive rate (TP) using two independent external validation cohorts. That is, the trained P-NET model correctly classified 73% of the primary tumors and 80% of the metastatic tumors. P-NET explicitly reduced the number of parameters for learning which have posed the challenges and successfully improved interpretability. They demonstrated that biologically informed deep neural networks such as P-NET they trained represent a novel approach to integrating cancer biology with machine learning by building mechanistic predictive models, providing a platform for biological discovery that may be widely applicable throughout cancer prediction and discovery tasks [124].

Case study 9. The systematic approach Machine Learning (ML) and Artificial Intelligence (AI) models use within the healthcare field

New models have been and created in ways that speed up the process of finding accurate predictions using new systematic approaches. Decision Trees, Random Forest, Support Vector Machines, and Artificial Neural Networks are a few examples of ML models that implement supervised learning techniques. Decision tree Models create a decision support tool that starts with a singular node and identifies the multitude of branching outcomes of one decision. The model then goes on to repeat finding outcomes of decisions using each previous product to identify a final product . Support Vector Machines are a type of classification model that uses supervised learning to identify features in two group situations by finding the largest marginal hyperplane to segregate the data, then sorts it out. Artificial Neural Networks are made up of an input layer, one or more hidden layers, and output layers. In these layers functional neurons in a singular layer are connected to every neuron in each level before or after . Some unsupervised learning models include the k-Means algorithm, Deep Belief Networks, and Convolutional Neural Networks. The most well known unsupervised learning model is the K-Means model that is used to identify the mean between groups of unlabeled data sets and create groups based on the mean. A Deep Belief Network (DBN) is a multi-layer network consisting of Indra-level connections useful for data retrieval that normally utilizes unsupervised learning and consists of many hidden layers tasked with feature detection and identifying correlations within the data. A Convolutional Neural Network (CNN) is a multilayer network that relies on feature recognition and identification and is useful for anomaly detection, image recognition, and identification [125]. Before these models may contribute to anything, there are certain steps that must be taken for the model to use the data provided. Data collection is the first stage, involving the collection of relevant data from various sources such as databases or sensors. The data consists of case scenarios organized and grouped according to specific criteria that are relevant to decision making. The second stage, Data pre-processing and cleaning is tasked with handling any missing values, outliers, noise, and inconsistencies. Data processing techniques may include data cleaning, normalization, feature scaling, encoding categorical variables, and dimensionality reduction. Feature engineering involves selecting, transforming, and creating new features from raw data to enhance the performance of Machine Learning models. This may include extracting meaningful features from raw data, combining multiple features, or generating new features using domain knowledge. Data modeling is a crucial part of the architecture, as it aims to build a model capable of learning patterns and relationships between input features and output labels. Model evaluation is the final stage. A post training model is sent to evaluation, it is evaluated using a validation set in order to assess its performance on unseen data. The best model is then chosen from among those that have been evaluated. After the final stage has been completed, the best model is ready to be used in a healthcare setting [126].

B. System Architecture

Data Collection

Refers to the process that carries the inputs necessary to understand the biological traits of each individual. During this initial phase, a wide range of individual data is gathered and stored, namely clinical, genetic, biomarker, and lifestyle data. [127]. Furthermore, biobanks hold significant importance as digital biorepositories, enabling the identification of reliable biomarkers for implementing precision medicine techniques [128].

Data Analytics

Allows the identification of anomalies or heterogeneity in data to pre-process them by filtering out inconsistent and incorrect data [129]. After that, the clinical data are standardized, which in turn fuses them further for analysis. Respectively, the most promising and popular technique at this stage is clustering [130]. This method classifies a dataset into subgroups of instances with similar components to each other by making use of high-throughput molecular technology [131]. Hence, this high-dimensional molecular data and the evolutionary progress of diseases can therefore be subdivided into more homogeneous groups of data belonging to patients suffering from complex diseases. Clustering techniques have been applied in the field of microbiomics for the amplification of hypervariable fragments of bacterial 16S rRNA genes, which are subsequenced before their extrapolation to taxonomic units [132]. This practice, hence allows labeling and understanding of drug interactions by bacteria in various human microbiomes. For example, digoxin is inactivated under the influence of some specific gut bacteria [132]; so getting the complete profile of the microbiome accelerates personalized and preventive drug treatment creation against side effects. This develops an interest in data analysis using the clustering technique. In this case, since the microbiome modifies a person's metabolome and epigenome, the clustering technique to capture classified microbiome profiles is of paramount importance in precision medicine [14]. This phase is critical since it determines the quality of the data and, therefore, the quality of the individual cure.

Model Development and Validation

Data mining is considered to be the extraction of useful information from complex, stratified databases. Big data mining, in particular, as a technique for bioinformatics, has shown to be a process beneficial for populations in a myriad of diseases [133]. Analysis of data is done in such a way that it identifies risks and opportunities, therefore unleashing the patterns and knowledge of genes and their relations with the natural evolution of disease to produce more pragmatic knowledge [134]. Hence, this novel process has impacted breast and ovarian cancer research in that it facilitates the discovery of BRCA1 and BRCA2 genes that affect the abrupt growth of these cancers such that physicians can order BRCA1 and BRCA2 tests based on family and personal history to reduce their chances of acquiring these cancers [135]. Having a more fundamental understanding of the data collected, several parallel models are developed which are declared accurate if they truly generalize and respond inclusively to the unknown data; if this standard is not met, these models are continuously trained until they show to meet the established parameters. Model validation in the field of precision medicine should take place by applying the existing rules regarding good data quality and openness with performance assessment within rigorous testing and appropriate licensing [136]. Once these models get validated, they are catered into clinical practice as part of a decision support systems.

Decision Support System

The Decision Support System (DSS) employs integrated data and models which functions to assist physicians in making decisions by supplying proposals grounded on an individual patient's unique clinical as well as personal history. Those systems have turned out to be a breakthrough since they offer structured and understandable information with access to medical literature while at the same time dealing with problems related to information overflow, and little human memory, hence cutting down on costs through better diagnoses and prognoses; they allow questions to be asked about aggregated patient data, and they provide warnings and suggestions about clinical decisions [137]. In the context of advancing genomic technologies in pediatric care, a DSS has been developed around neonatal genetic screening for pediatric cancer. This system integrates clinical, genetic, and epidemiological data to assist physicians in the interpretation of genomic screening results to improve the early detection and treatment of patients prone to contracting this disease [138].

Patient Participation

Patient concerns and uncertainties are addressed to avoid problems related to resistance to the adoption of these technologies [139]. As a result, detailed ethical terms and conditions policies should be established to provide patients with confidence in the use and privacy of their personal information and the results of subsequent analyses. Moreover, achieving the above will prompt patients to be actively involved in these personalized therapies to decentralize and versatilize these technologies. To illustrate, Mikhaylova and Thornton demonstrated that the performance of PrediXcan (a popular statistical method in precision medicine) differs according to the geographic origin of the population, with the rate being lower in populations of African descent compared to other populations of European descent [140]. Indeed, if a model is developed and trained on a given range of data, its application is not likely to be transferable to other populations. Some of the factors that make the universal application of these

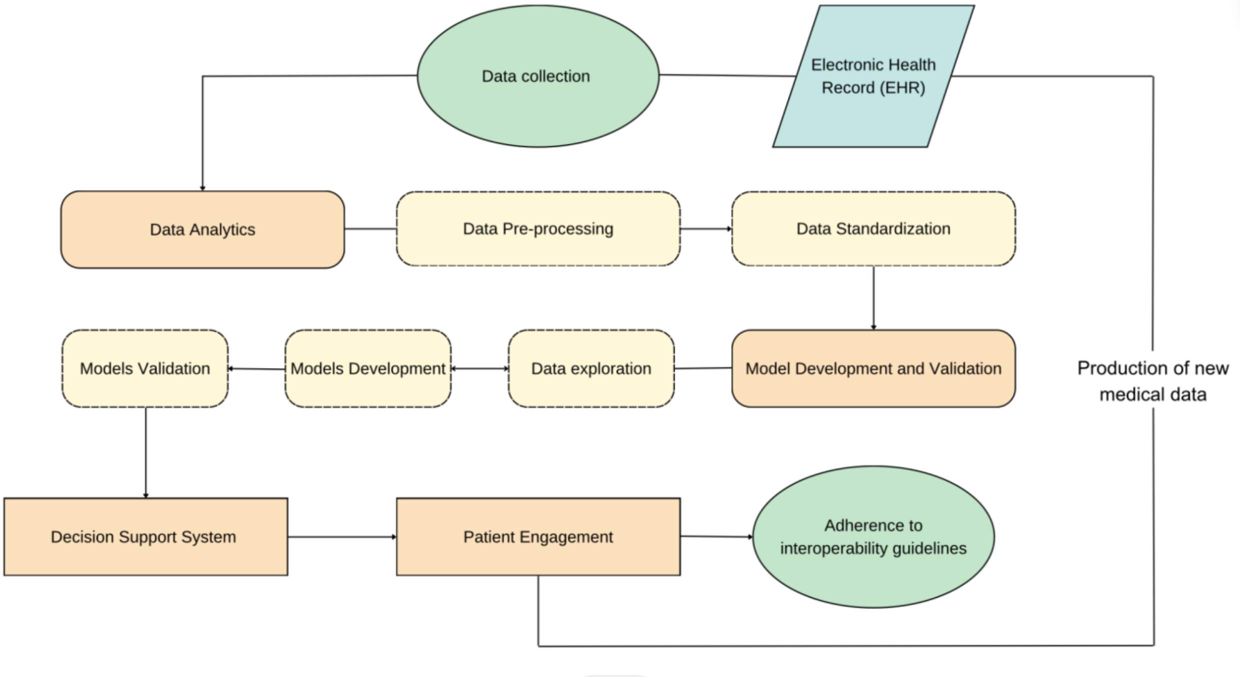

Figure 4. System architecture for machine learning models

models impossible are differences in allele frequency, linkage disequilibrium, and genetic admixture [141]. Nevertheless, precision medicine interoperability emerges as a possible solution to this concern.

System Interoperability

Interoperability encompasses making various software and data storage systems share information consistently, effectively, and responsibly [142]. To ensure that precision medicine maximizes its results for each patient, global intersectional data-sharing guidelines between different institutions promoting these types of computational advances are required as a precondition. However, there is still a noticeable fragmentation of policy among leading data warehousing organizations [143]. The four types of interoperability that should prevail in the data-sharing ecosystem are technical interoperability (data exchange between computer systems), syntactic interoperability (structure and format of the exchanged data), semantic interoperability (meaning of the shared data), and organizational interoperability (adoption of policies and guidelines) [144]. Therefore, the government plays an essential role in promoting data security standards, surveillance systems, and simultaneous multidisciplinary collaboration with the public and private sectors [145]. This is the only way to optimize and democratize precision medicine for the global population (Figure 4).

C. Ethical Considerations

Informed Consent

One such consideration is the act of garnering informed consent from patients. The ethical dilemma here would be ensuring the full understanding of patients and taking careful measures not to pressure them into giving consent. Informed consent is made up of three distinct constructs [146]. The first of which states that all information and risks involved in the study must be explained to its participants, regardless of the effect it may have on their willingness to contribute. The second construct says that participants must be fully able to understand the risks which come with participation in the study, and the information which has been divulged to them by researchers. The third construction demands that the participant has the option of voluntary consent which refers to the ability of an individual to join or leave a study. Although informed consent is applicable in most cases, in precision medicine oriented settings, which use tools such as big data, it becomes harder for participants to fully understand the scope of the data presented to them. This is especially prevalent in pharmacogenetic (PGx) testing. PGx testing examines the likelihood of an individual to display a negative or positive response to a drug, allowing for drug selection to be more tailored to the beneficiary [147]. As such complex tests can be difficult to explain to patients, it makes it harder for researchers to gain informed consent. This poses a new challenge for researchers who must find new, more effective ways of gaining consent to continue their research in an ethical manner.

Data-Driven Doctor-Patient Relationship

In the physician-patient relationship there are two roles, the sick patient and the physician [148]. In a setting which does not use precision medicine methods, the patient would normally come to the physician with a problem. The physician, on their part, would do their best to remedy the issue be it through diagnosis, referral, disconfirmation, or various medical methods [148]. When the doctor-patient relationship is used in precision medicine, though, the sick role becomes more ambiguous. As precision medicine relies on analyzing risk factors and trying to target diseases before they occur, patients are no longer classified as simply sick, but bear more similarity to patients-in-waiting. This is a state of uncertainty when it is unclear if or when an individual will develop a condition. This leads to increasing levels of anxiety in which individuals are unsure if they are healthy or sick [148]. As precision medicine also utilizes data about an individual's environment and lifestyle, a patient's anonymity may be compromised. Furthermore, the role of a patient may become more similar to that of a research participant, contributing to an ethical wrong on the patient's behalf. Meanwhile, physicians will begin to lose control of what data is revealed about their patients, leading to distrust between patients and doctors [148]. As the medical field becomes more data oriented, it poses the risk of dehumanizing patients, distilling them into data on a screen.

Clinical Benefit & Evidence for new Precision Medicine tests

In clinical settings, it takes time for evidence to be used to produce new precision medicine tests [149]. When deciding how to produce new tests, clinicians determine if the test will sufficiently benefit the patients under their care, posing the third ethical challenge of implementing precision medicine. Other factors used to determine sufficiency include the estimated benefit, pre-existing alternative treatments, and potential negative effects. In situations where evidence is limited but need for a new treatment is high, a consideration would be to implement clinical intervention and introduce the test in some clinical settings while normal treatment is offered in other institutions. Furthermore, the implementation of precision medicine includes its own benefits and risks, by thoughtfully addressing any ethical concerns, immorality will be effectively mitigated.

D. Challenges and Limitations

Models in machine learning are beginning to unleash their potential in precision medicine. These developments have outperformed the classical techniques of improved prediction accuracy and efficiently made successful classifications. Not only that, they also enable personalized treatments by merging algorithms trained with huge, complicated datasets that were impossible to be attained by the traditional ways. They have been drawing attention, especially in complex and heterogeneous diseases like cardiovascular diseases and neurodegenerative diseases that require revolutionary approaches.

There are several limitations to data-driven models: dependence on data, no transparency, higher computational cost, which also opens up security vulnerabilities. Since they learn from the data they are trained for, they inherit biases if they exist in the data itself. This creates a situation with potentially inaccurate or unfair results for groups underrepresented in the training data. On the other hand, in some such models, it's difficult to understand the reasons behind their predictions; this can make it hard to address errors or bias. Depending on the complexity, machine learning models can be costly because they could require a lot of computing resources. Also, they are prone to malicious attacks and can pose risks to data privacy because they often involve sensitive data, and security breaches or leaks can have serious consequences.

One of the most traditional problems throughout every field in machine learning is overfitting, which alludes to the prediction errors that occur for data sets lacking training. However, besides overfitting, there are several issues that are slightly different depending on certain domains. In drug toxicity prediction, model accuracy and generalization quality depend heavily on training data quality and diversity, while issues like class imbalance, validation accuracy and noisy labels remain big obstacles. In implementing big data approaches within cardiovascular medicine, several issues remain: data quality; integration of data from different healthcare systems; privacy and ethical concerns in the use of personal health data, and cause-result relationship determination from observational and retrospective data. Moreover, interpretability and transparency in the big data analytics models are essential factors to foster clinician trust and their adoption in clinical practice. AI in targeted drug delivery in cancer therapy remains a difficulty due to data bias in respect of ethnicity, data acquisition by independent institutes, code sharing for reproducibility, and integration across multiple sources. The complexity of cancer biology and the dynamic nature of tumor evolution increases the challenges even more.

The basic requirements for the safety and efficiency of AI-guided treatment strategies will have to be matched by rigorous validation and clinical trials. Data isolation, reproducibility, and data sparseness, as well as the complexity of the brain and the multifactorial nature of these diseases, are providing significant obstacles for computational techniques in neurodegenerative diseases. Collaboration between computational scientists, neurologists, and biologists is an important aspect to effectively translate these techniques into clinical practice and needs to be promoted in further studies.

Although the clustering technique facilitates the stratification of multimodal medical data into more homogeneous groups, a challenge in its implementation, "the curse of dimensionality", remains latent. This obstacle alludes to the creation of blind spots (regions with missing values and samples) during big data clustering [150]. This occurs due to the disproportionate increase in the dimensionality of the data (features and types), since if the algorithm detects a new category in the data set, by nature, it will generate a new classification with a vague variety of data to train itself to perform timely and accurately in clinical practices. As for the big data mining technique, there is an inherent challenge that limits the capability of this method. It is the heterogeneity in the format of previously classified databases since extrapolating mixed data (email attachments, images, pdf documents, medical records, X-rays, voice mails, graphics, video, audio, etc) to a standard format requires sophisticated software that has not yet been explored. [151]. This entails deficiencies in the development of computational models based on big data, affecting the increase in the speed of response in clinical situations that require immediate analysis without compromising the quality of the results. Regarding the DSSs, evidence has shown them to be an excellent support in advancing the quality and specificity of treatments. However, the current human-computer interface (HCI) paradigm lacks a responsive and interactive digital environment that is adapted to specific clinical scenarios. [152]. A gap that constrains the prevention of errors of omission and commission in the framework of precision medicine is the development of HCIs that allow clinicians to be reminded of aspects they have overlooked, including requested as well as unrequested suggestions, so that these systems are independent to intercede when necessary, but without going to the extent of being intrusive. The need for inclusivity practices was also identified as a diligent and concomitant factor in precision medicine, as this will trigger equitable and beneficial services for any individual. Analogously, some healthcare system firms still refuse to change their traditional paper-based system records to electronic health systems [153], making this resistance impossible to fully embrace the principles of interoperability, as these hand-written data are being excluded. In the same way, despite efforts to achieve semantic interoperability in the healthcare domain, a myriad of incompatible ontologies and terminologies are still emerging [154]. This significantly interferes with the smooth exchange of data, as multiple meanings are being attributed to similar clinical expressions.

The preservation of anonymity is crucial for the integration of precision medicine as the amount of data collected for research increases. Data can be gained from electronic medical records, lifestyle choices, or other devices used to monitor health. This expansion of information gathering can lead to a loss of control over where data goes. These new technologies had allowed it to become easier for doctors and researchers to access previously off-limit information, resulting in less patient anonymity. When gathering data it has become harder to do so ethically with the challenges posed by informed consent. Because informed consent requires patients to understand all aspects of a test or treatment, data becomes more difficult to obtain. This leads to using alternative forms of consent which can be more invasive into a patient’s privacy and medical health. Finally, it is important to consider the effects that precision medicine has on the doctor-patient relationship. As precision medicine becomes more prominent and based in data analysis, it runs the risk of focusing less on patient care and more on analytics. This can negatively affect patients by increasing the amount of data they must understand, contributing to stress, and endangering the sense of safety and privacy commonly shared by doctors and patients. At this point, the question moves on to seeking another solution aside from ide-identification, which can be easily re-identified in a variety of ways. We believe that further discovery to address the data privacy issues should be pursued in the future.

2. Conclusion

Utilizing data-driven techniques has become critical for improving precision medicine procedures. This review examines the obstacles and solutions of incorporating computational approaches into healthcare settings. We discuss the importance of modeling in healthcare, highlighting the limitations of traditional techniques, and the potential benefits of data-driven methodology. Data-driven approaches such as ML and deep learning algorithms are significant in various scopes of precision medicine including drug discovery, cardiovascular diseases, and neurodegenerative diseases. Researchers utilize a variety of data types to create ML models for precision medicine, which adhere to a structured system architecture from data collection to patient interaction in personalized treatment. Aspects such as the use of Decision Support Systems, patient engagement, and worldwide data-sharing standards illustrate the need for collaboration in improving global health outcomes [136]. Implementing data-driven models or computer-generated decisions in a healthcare setting is a complex strategy that raises numerous ethical concerns such as informed consent, physician-patient relationships, and patient autonomy. While medical science becomes more data-driven, it drives patients to the periphery and makes them no more than statistics on a computer screen [149]. Despite the positive advances in the field of AI and ML, the obstacles including the lack of interpretation capability, vulnerable data security, high dependency on data quality, etc, still inhibit possibilities for progress. The sample size of the studies considered within the applications of precision medicine is limited; therefore, the findings and results represent a partial view of the present field of study, as other fields of application in infectious, respiratory, and endocrine diseases were not explored. These recent advancements in data-based modeling will only give rise to more accurate and personalized treatments. Our proposed system architecture and machine learning algorithms have been evaluated only in select areas (e.g., medication toxicity prediction, cardiovascular medicine, cancer therapy, neurodegenerative illnesses) and on a small number of data sets. This reduces the theoretical generalizability of our findings. Also, potential biases in the data and research used to train and validate the ML models can have an influence on their robustness and application to larger, more varied populations. The perspective-based methodology used to analyze external studies implies the existence of biased and subjective interpretations throughout the research, adding to the fact that our personal experience and knowledge may have influenced scientific observations and conclusions. However, this obstacle is universal due to the open-ended nature of this research format [155]. In particular, the system architecture was designed based on logical and hierarchical criteria, since no other studies associated with the integration flow of a computational advance within precision medicine were found, therefore, the proposed system architecture model is subjective and its applicability may differ from other scenarios and practices in the field. Furthermore, another limitation in the study is that we brought up the creation process of ML models, rather than drilling down into how ML models are developed fundamentally since we have adopted a focus toward explaining ML models' outcomes, applications, advantages, and disadvantages. Expanding research in precision medicine will improve treatment accuracy and personalized care, despite the inherent limitations and diverse application of computational advancements in different situations. Based on current status and development that has been made in recent decades elucidated in our review, exploring in-depth integration of machine learning for precision medicine and constructing system modeling methods for different algorithms for real world applications are suggested in later work. As much as data-driven methods are illuminated, real-world data collection and storage ought to be improved, such as complete utilization of EHRs by addressing their noises, sparseness, and inconsistency that have prevented them from making a significant contribution as one of the most abundant patient health data resources. Particularly for implementation of machine learning models in clinical settings, robust validation criteria and parameters based on professional judgment and real world statistical model performance should necessarily be defined. Model creation for prediction and diagnosis appears to be focused rather than its evaluation. Few published research went beyond the empirical reports and described patient tests conducted in real world operation but even those studies did not provide detailed specifics about their processes for models. It remains problematic especially due to the critical nature of the health industry that mistakes could put patients into death. Thus, in order to improve reliability, a great amount of attention should be focused on continuous model evaluation and monitoring.

Author contributions

K.G.B., K.D., S.V.: conceptualization. N.J., S.R., S.K., R.K., J.C., N.M., K.G.B., K.D, S.V.: methodology, writing (original draft, review, editing), resources. S.K., R.K., N.M.: Case Studies. S.K., R.K., J.C., S.R., N.J.,: Visualization

Competing financial interests

The authors declare no competing financial interests.

References

1. T. I. Sørensen, "Which patients may be harmed by good treatments?," *The Lancet*, vol. 348, no. 9024, pp. 351–352, Aug. 1996, doi: https://doi.org/10.1016/s0140-6736(05)64988-4.
2. N. T. Longford and J. A. Nelder, "Statistics versus statistical science in the regulatory process," Statistics in medicine, vol. 18, no. 17–18, pp. 2311–2320, Sep. 1999, doi: https://doi.org/10.1002/(sici)1097-0258(19990915/30)18:17/18%3C2311::aid-sim257%3E3.0.co;2-t.
3. R. L. Kravitz, N. Duan, and J. Braslow, "Evidence-Based Medicine, Heterogeneity of Treatment Effects, and the Trouble with Averages," *The Milbank quarterly*, vol. 82, no. 4, pp. 661–687, Dec. 2004, doi: https://doi.org/10.1111/j.0887-378x.2004.00327.x.
4. "FACT SHEET: President Obama's Precision Medicine Initiative," whitehouse.gov, Jan. 30, 2015. https://obamawhitehouse.archives.gov/the-press-office/2015/01/30/fact-sheet-president-obama-s-precision-medicine-initiative (accessed Jun. 27, 2024).
5. "Google Books," *Google.com*, 2019. https://www.google.com/books/edition/MEDICAL_AND_HEALTH_SCIENCES_Volume_VIII/t9bCDAAAQBAJ?hl=en&gbpv=1&pg=PA53&printsec=frontcover (accessed Jun. 27, 2024).
6. C. J. Phillips, "Precision Medicine and its Imprecise History," *Harvard Data Science Review*, Jan. 2020, doi: https://doi.org/10.1162/99608f92.3e85b56a.
7. J. McCarthy, M. L. Minsky, N. Rochester, and C. E. Shannon, "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955," *The AI magazine/AI magazine*, vol. 27, no. 4, pp. 12–12, Dec. 2006, doi: https://doi.org/10.1609/aimag.v27i4.1904.
8. A. L. Samuel, "Some studies in machine learning using the game of checkers," *IBM journal of research and development*, vol. 44, no. 1.2, pp. 206–226, Jan. 2000, doi: https://doi.org/10.1147/rd.441.0206.
9. R. Dechter, "Learning While Searching in Constraint-Satisfaction-Problems - AAAI," *AAAI*, Oct. 16, 2023. https://aaai.org/papers/00178-aaai86-029-learning-while-searching-in-constraint-satisfaction-problems/ (accessed Jun. 27, 2024).
10. Y. LeCun *et al.*, "Backpropagation Applied to Handwritten Zip Code Recognition," *Neural computation*, vol. 1, no. 4, pp. 541–551, Dec. 1989, doi: https://doi.org/10.1162/neco.1989.1.4.541.
11. "The Medical Futurist," *The Medical Futurist*, 2024. https://medicalfuturist.com/fda-approved-ai-based-algorithms/ (accessed Jun. 27, 2024).
12. B. Mesko, "The role of artificial intelligence in precision medicine," *Expert Review of Precision Medicine and Drug Development*, vol. 2, no. 5, pp. 239–241, Sep. 2017, doi:

https://doi.org/10.1080/23808993.2017.1380516.

13. K. B. Johnson *et al.*, “Precision Medicine, AI, and the Future of Personalized Health Care,” *Clinical and translational science*, vol. 14, no. 1, pp. 86–93, Oct. 2020, doi: https://doi.org/10.1111/cts.12884.
14. Nardeep Naithani, S. Sinha, P. Misra, B. Vasudevan, and R. Sahu, “Precision medicine: Concept and tools,” *Medical Journal Armed Forces India/MJAFI*, vol. 77, no. 3, pp. 249–257, Jul. 2021, doi: https://doi.org/10.1016/j.mjafi.2021.06.021.
15. L. Zhang *et al.*, “Applications of Machine Learning Methods in Drug Toxicity Prediction,” *Current Topics in Medicinal Chemistry*, vol. 18, no. 12, pp. 987–997, Sep. 2018, doi: https://doi.org/10.2174/1568026618666180727152557.
16. Ahmet Erdemir *et al.*, “Credible practice of modeling and simulation in healthcare: ten rules from a multidisciplinary perspective,” *Journal of translational medicine*, vol. 18, no. 1, Sep. 2020, doi: https://doi.org/10.1186/s12967-020-02540-4.
17. H. J. Pandya et al., “Label-free electrical sensing of bacteria in eye wash samples: A step towards point-of-care detection of pathogens in patients with infectious keratitis,” Biosensors and Bioelectronics, vol. 91, pp. 32–39, May 2017, doi: https://doi.org/10.1016/j.bios.2016.12.035.
18. S. Odinotski et al., “A Conductive Hydrogel-Based Microneedle Platform for Real-Time pH Measurement in Live Animals,” Small, vol. 18, no. 45, Sep. 2022, doi: https://doi.org/10.1002/smll.202200201.
19. Center, “Computational Modeling Studies in Medical Device Submissions,” *U.S. Food and Drug Administration*, 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/reporting-computational-modeling-studies-medical-device-submissions (accessed Jun. 27, 2024).
20. K.-K. Mak and Mallikarjuna Rao Pichika, “Artificial intelligence in drug development: present status and future prospects,” *Drug discovery today*, vol. 24, no. 3, pp. 773–780, Mar. 2019, doi: https://doi.org/10.1016/j.drudis.2018.11.014.
21. R. Gupta, D. Srivastava, M. Sahu, S. Tiwari, R. K. Ambasta, and P. Kumar, “Artificial intelligence to deep learning: machine intelligence approach for drug discovery,” *Molecular diversity*, vol. 25, no. 3, pp. 1315–1360, Apr. 2021, doi: https://doi.org/10.1007/s11030-021-10217-3.
22. Peyman GhavamiNejad, Amin GhavamiNejad, H. Zheng, K. Dhingra, M. Samarikhalaj, and Mahla Poudineh, “A Conductive Hydrogel Microneedle-Based Assay Integrating PEDOT:PSS and Ag-Pt Nanoparticles for Real-Time, Enzyme-Less, and Electrochemical Sensing of Glucose,” Advanced Healthcare Materials, vol. 12, no. 1, Oct. 2022, doi: https://doi.org/10.1002/adhm.202202362.
23. R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, “An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review,” *Information*, vol. 15, no. 4, pp. 235–235, Apr. 2024, doi: https://doi.org/10.3390/info15040235.
24. Editor, “A Handy Guide to Random Forests for Big Biomedical Data | Editage Blog,” *Educational Articles For Researchers, Students And Authors - Editage Blog*, Nov. 2023. https://www.editage.com/blog/random-forests-for-big-biomedical-data/#:~:text=Random%20Forests%20have%20found%20extensive,%2C%20clinical%2C%20or%20omics%20data. (accessed Jun. 27, 2024).
25. Y. Wu and G. Wang, “Machine Learning Based Toxicity Prediction: From Chemical Structural Description to Transcriptome Analysis,” *International journal of molecular sciences*, vol. 19, no. 8, pp. 2358–2358, Aug. 2018, doi: https://doi.org/10.3390/ijms19082358.
26. Ruthwik Guntupalli, Saloni Verma and Karan Dhingra (2024) “Impact of Healthcare Digitization: Systems Approach for Integrating Biosensor Devices and Electronic Health with Artificial Intelligence”, American Scientific Research Journal for Engineering, Technology, and Sciences,

98(1), pp. 246–257. Available at: https://asrjetsjournal.org/index.php/American_Scientific_Journal/article/view/10786 (Accessed: 28 August 2024).

27. J. A. Leopold, B. A. Maron, and J. Loscalzo, “The application of big data to cardiovascular disease: paths to precision medicine,” *Journal of Clinical Investigation*, vol. 130, no. 1, pp. 29–38, Jan. 2020, doi: https://doi.org/10.1172/jci129203.
28. I. Hussain and M. B. Nazir, "Precision Medicine: AI and Machine Learning Advancements in Neurological and Cardiac Health," Revista Espanola de Documentacion Cientifica, vol. 18, no. 02, pp. 150-179, April. 2024
29. M. Safavieh et al., “Paper microchip with a graphene-modified silver nano-composite electrode for electrical sensing of microbial pathogens,” Nanoscale, vol. 9, no. 5, pp. 1852–1861, 2017, doi: https://doi.org/10.1039/c6nr06417e.
30. C. Strafella *et al.*, “Application of Precision Medicine in Neurodegenerative Diseases,” *Frontiers in Neurology*, vol. 9, Aug. 2018, doi: https://doi.org/10.3389/fneur.2018.00701.
31. Gupte, P.; Dhingra, K.; Saloni , V. Precision Gene Editing Strategies With CRISPR-Cas9 for Advancing Cancer Immunotherapy and Alzheimer’s Disease. J. Knowl. Learn. Sci. Technol. 2024, 3 (4), 11-21. https://doi.org/10.60087/jklst.v3.n4.p11.
32. B. Yin *et al.*, “Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype,” *Bioinformatics*, vol. 35, no. 14, pp. i538–i547, Jul. 2019, doi: https://doi.org/10.1093/bioinformatics/btz369.
33. K. P. Das and C. J, “Nanoparticles and convergence of artificial intelligence for targeted drug delivery for cancer therapy: Current progress and challenges,” *Frontiers in Medical Technology*, vol. 4, Jan. 2023, doi: https://doi.org/10.3389/fmedt.2022.1067144.
34. H. J. Pandya et al., “A microfluidic platform for drug screening in a 3D cancer microenvironment,” Biosensors and Bioelectronics, vol. 94, pp. 632–642, Aug. 2017, doi: https://doi.org/10.1016/j.bios.2017.03.054.
35. B. Bhinder, C. Gilvary, N. S. Madhukar, and O. Elemento, “Artificial Intelligence in Cancer Research and Precision Medicine,” *Cancer Discovery*, vol. 11, no. 4, pp. 900–915, Apr. 2021, doi: https://doi.org/10.1158/2159-8290.cd-21-0090.
36. L. E. Kiss, István Kövesdi, and József Rábai, “An improved design of fluorophilic molecules: prediction of the ln P fluorous partition coefficient, fluorophilicity, using 3D QSAR descriptors and neural networks,” *Journal of fluorine chemistry*, vol. 108, no. 1, pp. 95–109, Mar. 2001, doi: https://doi.org/10.1016/s0022-1139(01)00342-6.
37. L. Zhang *et al.*, “Applications of Machine Learning Methods in Drug Toxicity Prediction,” *Current Topics in Medicinal Chemistry*, vol. 18, no. 12, pp. 987–997, Sep. 2018, doi: https://doi.org/10.2174/1568026618666180727152557.
38. J. Klekota and F. P. Roth, “Chemical substructures that enrich for biological activity,” *Computer applications in the biosciences*, vol. 24, no. 21, pp. 2518–2525, Nov. 2008, doi: https://doi.org/10.1093/bioinformatics/btn479.
39. L. H. Hall and L. B. Kier, “Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological, and Valence State Information,” *Journal of Chemical Information and Computer Sciences*, vol. 35, no. 6, pp. 1039–1045, Nov. 1995, doi: https://doi.org/10.1021/ci00028a014.
40. D. Rogers and M. Hahn, “Extended-Connectivity Fingerprints,” *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 742–754, Apr. 2010, doi: https://doi.org/10.1021/ci100050t.
41. M. W. H. Wang, J. M. Goodman, and T. E. H. Allen, “Machine Learning in Predictive Toxicology: Recent Applications and Future Directions for Classification Models,” *Chemical Research in*

Toxicology, vol. 34, no. 2, Dec. 2020, doi: https://doi.org/10.1021/acs.chemrestox.0c00316.

42. C. Hansch, "Quantitative approach to biochemical structure-activity relationships," *Accounts of Chemical Research*, vol. 2, no. 8, pp. 232–239, Aug. 1969, doi: https://doi.org/10.1021/ar50020a002.
43. S. P. Bradbury, "Predicting Modes of Toxic Action from Chemical Structure: An Overview," *SAR and QSAR in Environmental Research*, vol. 2, no. 1–2, pp. 89–104, Apr. 1994, doi: https://doi.org/10.1080/10629369408028842.
44. M. T. D. Cronin and J. C. Dearden, "QSAR in Toxicology. 1. Prediction of Aquatic Toxicity," *Quantitative Structure-Activity Relationships*, vol. 14, no. 1, pp. 1–7, 1995, doi: https://doi.org/10.1002/qsar.19950140102.
45. D. Dix, K. A. Houck, M. J. Martin, A. M. Richard, R. Woodrow Setzer, and R. J. Kavlock, "The ToxCast Program for Prioritizing Toxicity Testing of Environmental Chemicals," *Toxicological Sciences*, vol. 95, no. 1, pp. 5–12, Jan. 2007, doi: https://doi.org/10.1093/toxsci/kfl103.
46. F. Cheng *et al.*, "admetSAR: A Comprehensive Source and Free Tool for Assessment of Chemical ADMET Properties," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 3099–3105, Nov. 2012, doi: https://doi.org/10.1021/ci300367a.
47. S. Kim, "Exploring Chemical Information in PubChem," *Current Protocols*, vol. 1, no. 8, Aug. 2021, doi: https://doi.org/10.1002/cpz1.217.
48. R. Benigni, C. L. Battistelli, C. Bossa, O. Tcheremenskaia, and P. Crettaz, "New perspectives in toxicological information management, and the role of ISSTOX databases in assessing chemical mutagenicity and carcinogenicity," *Mutagenesis*, vol. 28, no. 4, pp. 401–409, Mar. 2013, doi: https://doi.org/10.1093/mutage/get016.
49. A. Gaulton *et al.*, "ChEMBL: a large-scale bioactivity database for drug discovery," *Nucleic Acids Research*, vol. 40, no. D1, pp. D1100–D1107, Sep. 2011, doi: https://doi.org/10.1093/nar/gkr777.
50. T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson, "BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities," *Nucleic Acids Research*, vol. 35, no. Database, pp. D198–D201, Jan. 2007, doi: https://doi.org/10.1093/nar/gkl999.
51. D. Wishart *et al.*, "T3DB: the toxic exposome database," *Nucleic Acids Research*, vol. 43, no. D1, pp. D928–D934, Nov. 2014, doi: https://doi.org/10.1093/nar/gku1004.
52. C. Knox *et al.*, "DrugBank 3.0: a comprehensive resource for 'Omics' research on drugs," *Nucleic Acids Research*, vol. 39, no. Database, pp. D1035–D1041, Nov. 2010, doi: https://doi.org/10.1093/nar/gkq1126.
53. J. H. Olker *et al.*, "The ECOTOXicology Knowledgebase: A Curated Database of Ecologically Relevant Toxicity Tests to Support Environmental Research and Risk Assessment," *Environmental Toxicology and Chemistry*, vol. 41, no. 6, pp. 1520–1539, Apr. 2022, doi: https://doi.org/10.1002/etc.5324.
54. U. Schmidt *et al.*, "SuperToxic: a comprehensive database of toxic compounds," *Nucleic Acids Research*, vol. 37, no. Database, pp. D295–D299, Jan. 2009, doi: https://doi.org/10.1093/nar/gkn850.
55. J. Liu, C. Xu, W. Yang, Y. Shu, W. Zheng, and F. Zhou, "Multiple similarly effective solutions exist for biomedical feature selection and classification problems," *Scientific Reports*, vol. 7, no. 1, Oct. 2017, doi: https://doi.org/10.1038/s41598-017-13184-8.
56. R. Liu, X. Yu, and A. Wallqvist, "Using Chemical-Induced Gene Expression in Cultured Human Cells to Predict Chemical Toxicity," *Chemical Research in Toxicology*, vol. 29, no. 11, pp. 1883–1893, Nov. 2016, doi: https://doi.org/10.1021/acs.chemrestox.6b00287.
57. J. Zhang, Nikolaos Berntenis, A. Roth, and M. Ebeling, "Data mining reveals a network of early-response genes as a consensus signature of drug-induced in vitro and in vivo toxicity," *The*

Pharmacogenomics Journal, vol. 14, no. 3, pp. 208–216, Jun. 2014, doi: https://doi.org/10.1038/tpj.2013.39.
58. Z. Isik, C. Baldow, C. V. Cannistraci, and M. Schroeder, "Drug target prioritization by perturbed gene expression and network information," *Scientific Reports*, vol. 5, no. 5, p. 17417, Nov. 2015, doi: https://doi.org/10.1038/srep17417.
59. M. Kotlyar, K. Fortney, and I. Jurisica, "Network-based characterization of drug-regulated genes, drug targets, and toxicity," *Methods*, vol. 57, no. 4, pp. 499–507, Aug. 2012, doi: https://doi.org/10.1016/j.ymeth.2012.06.003.
60. J. Yamane *et al.*, "Prediction of developmental chemical toxicity based on gene networks of human embryonic stem cells," *Nucleic Acids Research*, vol. 44, no. 12, pp. 5515–5528, May 2016, doi: https://doi.org/10.1093/nar/gkw450.
61. A. Mayr, G. Klambauer, T. Unterthiner, and S. Hochreiter, "DeepTox: Toxicity Prediction using Deep Learning," *Frontiers in Environmental Science*, vol. 3, no. 3, Feb. 2016, doi: https://doi.org/10.3389/fenvs.2015.00080.
62. M. D. Zeiler and R. Fergus, "Visualizing and Understanding Convolutional Networks," *Computer Vision – ECCV 2014*, vol. 8689, pp. 818–833, 2014, doi: https://doi.org/10.1007/978-3-319-10590-1_53.
63. J. Donahue *et al.*, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *openaccess.thecvf.com*, 2015. https://openaccess.thecvf.com/content_cvpr_2015/html/Donahue_Long-Term_Recurrent_Convolutional_2015_CVPR_paper.html
64. C. Angermueller, T. Pärnamaa, L. Parts, and O. Stegle, "Deep learning for computational biology," *Molecular Systems Biology*, vol. 12, no. 7, p. 878, Jul. 2016, doi: https://doi.org/10.15252/msb.20156651.
65. A. Maunz, M. Gütlein, M. Rautenberg, D. Vorgrimmler, D. Gebele, and C. Helma, "lazar: a modular predictive toxicology framework," *Frontiers in Pharmacology*, vol. 4, 2013, doi: https://doi.org/10.3389/fphar.2013.00038.
66. Natalja Fjodorova *et al.*, "Quantitative and qualitative models for carcinogenicity prediction for non-congeneric chemicals using CP ANN method for regulatory uses," *Molecular Diversity*, vol. 14, no. 3, pp. 581–594, Aug. 2009, doi: https://doi.org/10.1007/s11030-009-9190-4.
67. K. Hansen *et al.*, "Benchmark Data Set for in Silico Prediction of Ames Mutagenicity," *Journal of Chemical Information and Modeling*, vol. 49, no. 9, pp. 2077–2081, Aug. 2009, doi: https://doi.org/10.1021/ci900161g.
68. C. Xu *et al.*, "In silico Prediction of Chemical Ames Mutagenicity," *Journal of Chemical Information and Modeling*, vol. 52, no. 11, pp. 2840–2847, Oct. 2012, doi: https://doi.org/10.1021/ci300400a.
69. S. Ekins, A. J. Williams, and J. J. Xu, "A Predictive Ligand-Based Bayesian Model for Human Drug-Induced Liver Injury," *Drug Metabolism and Disposition*, vol. 38, no. 12, pp. 2302–2308, Sep. 2010, doi: https://doi.org/10.1124/dmd.110.035113.
70. C. Zhang, F. Cheng, W. Li, G. Liu, P. W. Lee, and Y. Tang, "In silicoPrediction of Drug Induced Liver Toxicity Using Substructure Pattern Recognition Method," *Molecular Informatics*, vol. 35, no. 3–4, pp. 136–144, Feb. 2016, doi: https://doi.org/10.1002/minf.201500055.
71. X. Li *et al.*, "In Silico Prediction of Chemical Acute Oral Toxicity Using Multi-Classification Methods," *Journal of Chemical Information and Modeling*, vol. 54, no. 4, pp. 1061–1069, Apr. 2014, doi: https://doi.org/10.1021/ci5000467.
72. C. Zhang *et al.*, "In silico prediction of hERG potassium channel blockage by chemical category approaches," *Toxicology Research*, vol. 5, no. 2, pp. 570–582, 2016, doi:

https://doi.org/10.1039/c5tx00294j.
73. E. Oikonomou *et al.*, “Environment and cardiovascular disease: rationale of the Corinthia study,” *Hellenic Journal of Cardiology*, vol. 57, no. 3, pp. 194–197, May 2016, doi: https://doi.org/10.1016/j.hjc.2016.06.001.
74. A. Bhatnagar, “Environmental Determinants of Cardiovascular Disease,” *Circulation Research*, vol. 121, no. 2, pp. 162–180, 2017, doi: https://doi.org/10.1161/circresaha.117.306458.
75. N. P. Pronk, P. L. Mabry, S. Bond, R. Arena, and M. A. Faghy, “Systems science approaches to cardiovascular disease prevention and management in the era of COVID-19: A Humpty-Dumpty dilemma?,” *Progress in Cardiovascular Diseases*, vol. 76, pp. 69–75, Jan. 2023, doi: https://doi.org/10.1016/j.pcad.2022.12.003.
76. S. Kathiresan and D. Srivastava, “Genetics of Human Cardiovascular Disease,” *Cell*, vol. 148, no. 6, pp. 1242–1257, Mar. 2012, doi: https://doi.org/10.1016/j.cell.2012.03.001.
77. the CARDIoGRAMplusC4D Consortium, “A comprehensive 1000 Genomes–based genome-wide association meta-analysis of coronary artery disease,” *Nature Genetics*, vol. 47, no. 10, pp. 1121–1130, Sep. 2015, doi: https://doi.org/10.1038/ng.3396.
78. S. Rani, Pankaj Bhambri, A. Kataria, A. Khang, and Arun Kumar Sivaraman, *Big Data, Cloud Computing and IoT*. CRC Press, 2023.
79. T. A. Kass-Hout, L. M. Stevens, and J. L. Hall, “American Heart Association Precision Medicine Platform,” *Circulation*, vol. 137, no. 7, pp. 647–649, Feb. 2018, doi: https://doi.org/10.1161/circulationaha.117.032041.
80. P.-A. Gourraud *et al.*, “Precision medicine in chronic disease management: The multiple sclerosis BioScreen,” *Annals of Neurology*, vol. 76, no. 5, pp. 633–642, Oct. 2014, doi: https://doi.org/10.1002/ana.24282.
81. H. Meshref, “Cardiovascular Disease Diagnosis: A Machine Learning Interpretation Approach,” *International Journal of Advanced Computer Science and Applications*, vol. 10, no. 12, 2019, doi: https://doi.org/10.14569/ijacsa.2019.0101236.
82. H. Benjamin, F. David, and S. Belcy, “Heart Disease Prediction Using Data Mining Techniques,” *2017 International Conference on Intelligent Computing and Control (IEC2)*, Jun. 2017, doi: https://doi.org/10.21917/ijsc.2018.0254.
83. C. S. Mayo, M. M. Matuszak, M. J. Schipper, S. Jolly, J. A. Hayman, and R. K. Ten Haken, “Big Data in Designing Clinical Trials: Opportunities and Challenges,” *Frontiers in Oncology*, vol. 7, Aug. 2017, doi: https://doi.org/10.3389/fonc.2017.00187.
84. C. Pislaru, M. M. Alashry, J. J. Thaden, P. A. Pellikka, M. Enriquez-Sarano, and S. V. Pislaru, “Intrinsic Wave Propagation of Myocardial Stretch, A New Tool to Evaluate Myocardial Stiffness: A Pilot Study in Patients with Aortic Stenosis and Mitral Regurgitation,” *Journal of the American Society of Echocardiography*, vol. 30, no. 11, pp. 1070–1080, Nov. 2017, doi: https://doi.org/10.1016/j.echo.2017.06.023.
85. R. Laaksonen *et al.*, “Plasma ceramides predict cardiovascular death in patients with stable coronary artery disease and acute coronary syndromes beyond LDL-cholesterol,” *European Heart Journal*, vol. 37, no. 25, pp. 1967–1976, Apr. 2016, doi: https://doi.org/10.1093/eurheartj/ehw148.
86. B. L. Filkins *et al.*, “Privacy and security in the era of digital health: what should translational researchers know and do about it?,” *American journal of translational research*, vol. 8, no. 3, pp. 1560–80, 2016, Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4859641/
87. A. Althubaiti, “Information Bias in Health research: definition, pitfalls, and Adjustment Methods,” *Journal of Multidisciplinary Healthcare*, vol. 9, no. 9, pp. 211–217, 2016.
88. H. Murakami *et al.*, “Accuracy of Wearable Devices for Estimating Total Energy Expenditure,” *JAMA Internal Medicine*, vol. 176, no. 5, p. 702, May 2016, doi:

https://doi.org/10.1001/jamainternmed.2016.0152.
89. J. M. Jakicic *et al.*, “Effect of Wearable Technology Combined With a Lifestyle Intervention on Long-Term Weight Loss,” *Obstetrical & Gynecological Survey*, vol. 72, no. 2, pp. 67–68, Feb. 2017, doi: https://doi.org/10.1097/01.ogx.0000512372.67520.49.
90. I. D. Dinov, “Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data,” *GigaScience*, vol. 5, no. 1, Feb. 2016, doi: https://doi.org/10.1186/s13742-016-0117-6.
91. R. Bellazzi, “Big Data and Biomedical Informatics: A Challenging Opportunity,” *Yearbook of Medical Informatics*, vol. 23, no. 01, pp. 08-13, Aug. 2014, doi: https://doi.org/10.15265/iy-2014-0024.
92. S. B. Scruggs *et al.*, “Harnessing the Heart of Big Data,” *Circulation research*, vol. 116, no. 7, pp. 1115–1119, Mar. 2015, doi: https://doi.org/10.1161/CIRCRESAHA.115.306013.
93. M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, “Identifying Personal Genomes by Surname Inference,” *Science*, vol. 339, no. 6117, pp. 321–324, Jan. 2013, doi: https://doi.org/10.1126/science.1229566.
94. C. J. Presley *et al.*, “Association of Broad-Based Genomic Sequencing With Survival Among Patients With Advanced Non–Small Cell Lung Cancer in the Community Oncology Setting,” *JAMA*, vol. 320, no. 5, pp. 469–477, Aug. 2018, doi: https://doi.org/10.1001/jama.2018.9824.
95. G. J. Hollands *et al.*, “The impact of communicating genetic risks of disease on risk-reducing health behaviour: systematic review with meta-analysis,” *BMJ*, vol. 352, p. i1102, Mar. 2016, doi: https://doi.org/10.1136/bmj.i1102.
96. P. Schnohr, P. Lange, J. Nyboe, M. Appleyard, and G. Jensen, “Gray hair, baldness, and wrinkles in relation to myocardial infarction: The Copenhagen City Heart Study,” *American Heart Journal*, vol. 130, no. 5, pp. 1003–1010, Nov. 1995, doi: https://doi.org/10.1016/0002-8703(95)90201-5.
97. S. M. Lesko, “A Case-Control Study of Baldness in Relation to Myocardial Infarction in Men,” *JAMA: The Journal of the American Medical Association*, vol. 269, no. 8, p. 998, Feb. 1993, doi: https://doi.org/10.1001/jama.1993.03500080046030.
98. E. S. Ford, D. S. Freedman, and T. Byers, “Baldness and Ischemic Heart Disease in a National Sample of Men,” *American Journal of Epidemiology*, vol. 143, no. 7, pp. 651–657, Apr. 1996, doi: https://doi.org/10.1093/oxfordjournals.aje.a008797.
99. A. Madani, R. Arnaout, M. Mofrad, and R. Arnaout, “Fast and accurate view classification of echocardiograms using deep learning,” *npj Digital Medicine*, vol. 1, no. 1, Mar. 2018, doi: https://doi.org/10.1038/s41746-017-0013-1.
100. C. Cano-Espinosa, G. González, G. R. Washko, M. Cazorla, and R. S. J. Estépar, “Automated Agatston Score Computation in non-ECG Gated CT Scans Using Deep Learning,” *Proceedings of SPIE--the International Society for Optical Engineering*, vol. 10574, p. 105742K, Feb. 2018, doi: https://doi.org/10.1117/12.2293681.
101. A. Termine *et al.*, “Multi-Layer Picture of Neurodegenerative Diseases: Lessons from the Use of Big Data through Artificial Intelligence,” *Journal of Personalized Medicine*, vol. 11, no. 4, p. 280, Apr. 2021, doi: https://doi.org/10.3390/jpm11040280.
102. N. Perakakis, A. Yazdani, G. E. Karniadakis, and C. Mantzoros, “Omics, big data and machine learning as tools to propel understanding of biological mechanisms and to discover novel diagnostics and therapeutics,” *Metabolism*, vol. 87, pp. A1–A9, Oct. 2018, doi: https://doi.org/10.1016/j.metabol.2018.08.002.
103. A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma, “Network Medicine in the Age of Biomedical Big Data,” *Frontiers in Genetics*, vol. 10, Apr. 2019, doi: https://doi.org/10.3389/fgene.2019.00294.

104. P. L. De Jager, H.-S. Yang, and D. A. Bennett, “Deconstructing and targeting the genomic architecture of human neurodegeneration,” *Nature Neuroscience*, vol. 21, no. 10, pp. 1310–1317, Sep. 2018, doi: https://doi.org/10.1038/s41593-018-0240-z.
105. R. Ranchal *et al.*, “Disrupting Healthcare Silos: Addressing Data Volume, Velocity and Variety with a Cloud-Native Healthcare Data Ingestion Service,” *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 11, pp. 1–1, 2020, doi: https://doi.org/10.1109/jbhi.2020.3001518.
106. A. Sethi, A. Sankaran, N. Panwar, S. Khare, and S. Mani, “DLPaper2Code: Auto-Generation of Code From Deep Learning Research Papers,” *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, Apr. 2018, doi: https://doi.org/10.1609/aaai.v32i1.12326.
107. X. Chen, H. Chen, S. Nan, X. Kong, H. Duan, and H. Zhu, “Dealing With Missing, Imbalanced, and Sparse Features During the Development of a Prediction Model for Sudden Death Using Emergency Medicine Data: Machine Learning Approach,” *JMIR Medical Informatics*, vol. 11, p. e38590, Jan. 2023, doi: https://doi.org/10.2196/38590.
108. W. van Rheenen *et al.*, “Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis,” *Nature Genetics*, vol. 48, no. 9, pp. 1043–1048, Jul. 2016, doi: https://doi.org/10.1038/ng.3622.
109. J. Schmidhuber, “Deep Learning in Neural Networks: an Overview,” *Neural Networks*, vol. 61, no. 61, pp. 85–117, Jan. 2015, doi: https://doi.org/10.1016/j.neunet.2014.09.003.
110. Project MinE ALS Sequencing Consortium, “Project MinE: Study Design and Pilot Analyses of a large-scale whole-genome Sequencing Study in Amyotrophic Lateral Sclerosis,” *European Journal of Human Genetics*, vol. 26, no. 10, pp. 1537–1546, Jun. 2018, doi: https://doi.org/10.1038/s41431-018-0177-4.
111. P. M. Visscher *et al.*, “10 Years of GWAS Discovery: Biology, Function, and Translation,” *The American Journal of Human Genetics*, vol. 101, no. 1, pp. 5–22, Jul. 2017, doi: https://doi.org/10.1016/j.ajhg.2017.06.005.
112. S. J. MacEachern and N. D. Forkert, “Machine learning for precision medicine,” *Genome*, vol. 64, no. 4, pp. 416–425, Apr. 2021, doi: https://doi.org/10.1139/gen-2020-0131.
113. L. Ding *et al.*, “Perspective on Oncogenic Processes at the End of the Beginning of Cancer Genomics,” *Cell*, vol. 173, no. 2, pp. 305-320.e10, Apr. 2018, doi: https://doi.org/10.1016/j.cell.2018.03.033.
114. N. Coudray *et al.*, “Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning,” *Nature Medicine*, vol. 24, no. 10, pp. 1559–1567, Sep. 2018, doi: https://doi.org/10.1038/s41591-018-0177-5.
115. Y. Sun *et al.*, “Identification of 12 cancer types through genome deep learning,” *Scientific Reports*, vol. 9, no. 1, Nov. 2019, doi: https://doi.org/10.1038/s41598-019-53989-3.
116. J. Yuan *et al.*, “Integrated Analysis of Genetic Ancestry and Genomic Alterations across Cancers,” *Cancer cell*, vol. 34, no. 4, pp. 549-560.e9, Oct. 2018, doi: https://doi.org/10.1016/j.ccell.2018.08.019.
117. C. Zang *et al.*, “High-Throughput Clinical Trial Emulation with Real World Data and Machine Learning: A Case Study of Drug Repurposing for Alzheimer’s Disease,” *medRxiv (Cold Spring Harbor Laboratory)*, Feb. 2022, doi: https://doi.org/10.1101/2022.01.31.22270132.
118. M. B. Makarious *et al.*, “Multi-modality machine learning predicting Parkinson’s disease,” *NPJ Parkinson’s disease*, vol. 8, no. 1, Apr. 2022, doi: https://doi.org/10.1038/s41531-022-00288-w.
119. J. Liu, P. Liu, M.-R. Lei, H.-W. Zhang, A.-L. You, and X.-R. Luan, “Readmission Risk Prediction Model for Patients with Chronic Heart Failure: A Systematic Review and Meta-Analysis,” *Iranian journal of public health*, vol. 51, no. 7, Jul. 2022, doi: https://doi.org/10.18502/ijph.v51i7.10082.

120. C.-Y. Guo, M.-Y. Wu, and H.-M. Cheng, “The Comprehensive Machine Learning Analytics for Heart Failure,” *International Journal of Environmental Research and Public Health*, vol. 18, no. 9, p. 4943, Jan. 2021, doi: https://doi.org/10.3390/ijerph18094943.
121. A. Lysenko, A. Sharma, K. A. Boroevich, and T. Tsunoda, “An integrative machine learning approach for prediction of toxicity-related drug safety,” *Life Science Alliance*, vol. 1, no. 6, p. e201800098, Nov. 2018, doi: https://doi.org/10.26508/lsa.201800098.
122. L. Pu, M. Naderi, T. Liu, H.-C. Wu, S. Mukhopadhyay, and M. Brylinski, “eToxPred: a machine learning-based approach to estimate the toxicity of drug candidates,” *BMC Pharmacology and Toxicology*, vol. 20, no. 1, Jan. 2019, doi: https://doi.org/10.1186/s40360-018-0282-6.
123. L. K. Vora, A. D. Gholap, K. Jetha, R. R. S. Thakur, H. K. Solanki, and V. P. Chavda, “Artificial Intelligence in Pharmaceutical Technology and Drug Delivery Design,” *Pharmaceutics*, vol. 15, no. 7, pp. 1916–1916, 2023, doi: https://doi.org/10.3390/pharmaceutics15071916.
124. H. A. Elmarakeby *et al.*, “Biologically informed deep neural network for prostate cancer discovery,” *Nature*, vol. 598, no. 7880, pp. 348–352, Oct. 2021, doi: https://doi.org/10.1038/s41586-021-03922-4.
125. H. Habehh and S. Gohel, “Machine Learning In Healthcare,” *Current Genomics*, vol. 22, no. 4, Jul. 2021, doi: https://doi.org/10.2174/1389202922666210705124359.
126. R. Guido, S. Ferrisi, D. Lofaro, and D. Conforti, “An Overview on the Advancements of Support Vector Machine Models in Healthcare Applications: A Review,” *Information*, vol. 15, no. 4, p. 235, Apr. 2024, doi: https://doi.org/10.3390/info15040235.
127. I. R. König, O. Fuchs, G. Hansen, E. von Mutius, and M. V. Kopp, “What is precision medicine?,” *European Respiratory Journal*, vol. 50, no. 4, p. 1700391, Oct. 2017, doi: https://doi.org/10.1183/13993003.00391-2017.
128. P. Korhan, S. Tercan Avcı, Y. Yılmaz, Y. Öztemur Islakoğlu, and N. Atabey, “Role of Biobanks for Cancer Research and Precision Medicine in Hepatocellular Carcinoma,” *Journal of Gastrointestinal Cancer*, vol. 52, no. 4, pp. 1232–1247, Dec. 2021, doi: https://doi.org/10.1007/s12029-021-00759-y.
129. S. García, S. Ramírez-Gallego, J. Luengo, J. M. Benítez, and F. Herrera, “Big data preprocessing: methods and prospects,” *Big Data Analytics*, vol. 1, no. 1, Nov. 2016, doi: https://doi.org/10.1186/s41044-016-0014-0.
130. E. Parimbelli, S. Marini, L. Sacchi, and R. Bellazzi, “Patient similarity for precision medicine: A systematic review,” *Journal of biomedical informatics*, vol. 83, pp. 87–96, Jul. 2018, doi: https://doi.org/10.1016/j.jbi.2018.06.001.
131. Å. Johansson *et al.*, “Precision medicine in complex diseases—Molecular subgrouping for improved prediction and treatment stratification,” *Journal of internal medicine*, vol. 294, no. 4, pp. 378–396, Apr. 2023, doi: https://doi.org/10.1111/joim.13640.
132. R. C. Wang and Z. Wang, “Precision Medicine: Disease Subtyping and Tailored Treatment,” *Cancers*, vol. 15, no. 15, pp. 3837–3837, Jul. 2023, doi: https://doi.org/10.3390/cancers15153837.
133. Kulkarni S, Dhingra K, Verma S., "Applications of CMUT Technology in Medical Diagnostics: From Photoacoustic to Ultrasonic Imaging", International Journal of Science and Research (IJSR), Volume 13 Issue 6, June 2024, pp. 1264-1269, https://www.ijsr.net/archive/v13i6/SR24619062609.pdf
134. Reinaldo Padilha França, A. Carolina, R. Arthur, and Yuzo Iano, “An overview of the impact of PACS as health informatics and technology e-health in healthcare management,” *Elsevier eBooks*, pp. 101–128, Jan. 2022, doi: https://doi.org/10.1016/b978-0-12-824410-4.00007-6.
135. E. A. Estape, Mary Helen Mays, and E. A. Sternke, “Translation in Data Mining to Advance Personalized Medicine for Health Equity,” *Intelligent information management*, vol. 08, no. 01, pp.

9–16, Jan. 2016, doi: https://doi.org/10.4236/iim.2016.81002.

136. C. B. Collin *et al.*, "Computational Models for Clinical Applications in Personalized Medicine—Guidelines and Recommendations for Data Integration and Model Validation," *Journal of Personalized Medicine*, vol. 12, no. 2, p. 166, Jan. 2022, doi: https://doi.org/10.3390/jpm12020166

137. P. D. Clayton and G. Hripcsak, "Decision support in healthcare," *International journal of bio-medical computing*, vol. 39, no. 1, pp. 59–66, Apr. 1995, doi: https://doi.org/10.1016/0020-7101(94)01080-k.

138. BalaSubramani Gattu Linga, Mohammed, T. Farrell, Hilal Al Rifai, Nader Al-Dewik, and M. Walid Qoronfleh, "Genomic Newborn Screening for Pediatric Cancer Predisposition Syndromes: A Holistic Approach," *Cancers*, vol. 16, no. 11, pp. 2017–2017, May 2024, doi: https://doi.org/10.3390/cancers16112017.

139. "Precision Medicine: From Science To Value | Health Affairs Journal," *Health Affairs*, 2018. https://www.healthaffairs.org/doi/full/10.1377/hlthaff.2017.1624 (accessed Jun. 28, 2024).

140. A. V. Mikhaylova and T. A. Thornton, "Accuracy of Gene Expression Prediction From Genotype Data With PrediXcan Varies Across and Within Continental Populations," *Frontiers in genetics*, vol. 10, Apr. 2019, doi: https://doi.org/10.3389/fgene.2019.00261.

141. J. N. Cooke, W. S. Bush, and D. C. Crawford, "Editorial: The Importance of Diversity in Precision Medicine Research," *Frontiers in genetics*, vol. 11, Aug. 2020, doi: https://doi.org/10.3389/fgene.2020.00875.

142. Heubusch K, "Interoperability: what it means, why it matters," *Journal of AHIMA*, vol. 77, no. 1, 2022, Accessed: Jun. 28, 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/16475733/

143. J. S. Beckmann and D. Lew, "Reconciling evidence-based medicine and precision medicine in the era of big data: challenges and opportunities," *Genome medicine*, vol. 8, no. 1, Dec. 2016, doi: https://doi.org/10.1186/s13073-016-0388-7.

144. M. Lehne, J. Sass, A. Essenwanger, J. Schepers, and S. Thun, "Why digital medicine depends on interoperability," *npj digital medicine*, vol. 2, no. 1, Aug. 2019, doi: https://doi.org/10.1038/s41746-019-0158-1.

145. L. P. Whitsel, J. Wilbanks, M. D. Huffman, and J. L. Hall, "The Role of Government in Precision Medicine, Precision Public Health and the Intersection With Healthy Living," *Progress in cardiovascular diseases*, vol. 62, no. 1, pp. 50–54, Jan. 2019, doi: https://doi.org/10.1016/j.pcad.2018.12.002.

146. F. K. Dankar, M. Gergely, B. Malin, Radja Badji, S. K. Dankar, and K. Shuaib, "Dynamic-informed consent: A potential solution for ethical dilemmas in population sequencing initiatives," *Computational and Structural Biotechnology Journal*, vol. 18, pp. 913–921, Jan. 2020, doi: https://doi.org/10.1016/j.csbj.2020.03.027.

147. S. B. Haga and R. Mills, "A review of consent practices and perspectives for pharmacogenetic testing," *Pharmacogenomics*, vol. 17, no. 14, pp. 1595–1605, Sep. 2016, doi: https://doi.org/10.2217/pgs-2016-0039.

148. G. Eyal *et al.*, "The physician–patient relationship in the age of precision medicine," *Genetics in Medicine*, vol. 21, no. 4, pp. 813–815, Sep. 2018, doi: https://doi.org/10.1038/s41436-018-0286-z.

149. D. M. Korngiebel, K. E. Thummel, and W. Burke, "Implementing Precision Medicine: The Ethical Challenges," *Trends in Pharmacological Sciences*, vol. 38, no. 1, pp. 8–14, Jan. 2017, doi: https://doi.org/10.1016/j.tips.2016.11.007.

150. V. Berisha *et al.*, "Digital medicine and the curse of dimensionality," *npj Digital Medicine*, vol. 4, no. 1, Oct. 2021, doi: https://doi.org/10.1038/s41746-021-00521-5.

151. J. K.U and J. M. David, "Issues, Challenges and Solutions : Big Data Mining," *Computer Science & Information Technology (CS & IT)*, Dec. 2014, doi: https://doi.org/10.5121/csit.2014.41311.

152. D. F. Sittig *et al.*, “Grand challenges in clinical decision support,” *Journal of Biomedical Informatics*, vol. 41, no. 2, pp. 387–392, Apr. 2008, doi: https://doi.org/10.1016/j.jbi.2007.09.003.

153. K. Rosati, C. Gordon, S. Owens, N. Plc, and M. Lamar, “THE QUEST FOR INTEROPERABLE ELECTRONIC HEALTH RECORDS: A Guide to Legal Issues in Establishing Health Information Networks.” Accessed: Jun. 30, 2024. [Online]. Available:https://www.crowell.com/a/web/9GkdjVhUuLs35YDxqnmk1X/4TtizN/2005-July_Quest_for_EHRs_Butler.pdf

154. O. Iroju, A. Soriyan, I. Gambo, and J. Olaleke, “Applications of swam robots View project Harnessing Technology for enhanced living View project Iroju Olaronke adeyemi college of education ondo Interoperability in Healthcare: Benefits, Challenges and Resolutions,” *International Journal of Innovation and Applied Studies*, vol. 3, no. 1, pp. 262–270, 2013, Available: http://bio3.giga.ulg.ac.be/record/wp-content/uploads/2022/11/Interoperability-in-Healthcare.pdf

155. S. Tenny, J. Brannan, and G. Brannan, “Qualitative study,” *National Library of Medicine*, Sep. 18, 2022. https://www.ncbi.nlm.nih.gov/books/NBK470395/