NRC Research Press

MINIREVIEW

Machine learning for precision medicine[1]

Sarah J. MacEachern and Nils D. Forkert

Abstract: Precision medicine is an emerging approach to clinical research and patient care that focuses on understanding and treating disease by integrating multi-modal or multi-omics data from an individual to make patient-tailored decisions. With the large and complex datasets generated using precision medicine diagnostic approaches, novel techniques to process and understand these complex data were needed. At the same time, computer science has progressed rapidly to develop techniques that enable the storage, processing, and analysis of these complex datasets, a feat that traditional statistics and early computing technologies could not accomplish. Machine learning, a branch of artificial intelligence, is a computer science methodology that aims to identify complex patterns in data that can be used to make predictions or classifications on new unseen data or for advanced exploratory data analysis. Machine learning analysis of precision medicine's multi-modal data allows for broad analysis of large datasets and ultimately a greater understanding of human health and disease. This review focuses on machine learning utilization for precision medicine's "big data", in the context of genetics, genomics, and beyond.

Key words: machine learning, deep learning, precision medicine.

Résumé : La médecine personnalisée est une approche émergente en matière de recherche clinique et de soins qui repose sur une connaissance et un traitement de la maladie qui intègre des données multimodales ou multi-omiques provenant de l'individu, ce qui permet de prendre des décisions personnalisées pour chaque patient. En raison des jeux de données à la fois massifs et complexes générés par une telle approche diagnostique, de nouvelles techniques pour traiter et comprendre ces données complexes sont nécessaires. L'apprentissage automatique, une branche de l'intelligence artificielle, est une méthodologie informatique qui vise à identifier des schémas complexes au sein des données, permettant ainsi de faire des prédictions ou des classifications sur de nouvelles données ou une exploration avancée de l'analyse de ces données. L'apprentissage automatique sur des données multimodales issues de la médecine personnalisée permet une analyse étendue de vastes ensembles de données et, ultimement, d'acquérir une meilleure compréhension de la santé humaine et des maladies. Cette synthèse se concentre sur l'emploi de l'apprentissage automatique sur les données massives obtenues en médecine personnalisée dans un contexte génétique, génomique, et plus encore. [Traduit par la Rédaction]

Mots-clés : apprentissage automatique, apprentissage profond, médecine personnalisée.

Introduction

Precision medicine, often also referred to as precision health, is a novel approach to understanding health and disease based on patient-individual data, including medical diagnoses, clinical phenotype (severity of disease, amount of functional impairment, etc.), biologic investigations including laboratory studies and imaging, and environmental, demographic, and lifestyle factors. Taken together, these data are considered multi-modal as they represent information from multiple domains. The evolution of precision medicine has been heavily influenced by the exponential increase in the amount of biologic data that can now be collected for each individual patient, in large part due to the advent of new technologies in the fields of medicine, genetics, metabolics, and imaging, among others. The amount and variety of diagnostic tests that can be performed produces an incredible amount of

Received 31 July 2020. Accepted 20 October 2020.

S.J. MacEachern. Department of Pediatrics, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.

N.D. Forkert. Department of Radiology, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada; Alberta Children's Hospital Research Institute, Cumming School of Medicine, University of Calgary, Calgary, AB, Canada.

Corresponding author: Nils D. Forkert (email: nils.forkert@ucalgary.ca).

[1]This Minireview is part of a collection on Genome Biology jointly published by *Genome* and *Biochemistry and Cell Biology*.

data that is challenging to understand and analyze for a single patient, and even more difficult in a dataset containing information from multiple patients. Thankfully, as more sophisticated diagnostic tests were being developed, the field of computer science likewise experienced an evolution, allowing for storage and analysis of these large volumes of data more efficiently than ever before. These two developments go hand in hand, with computer science methodologies making use of the large volumes of deep data collected in the health care system, allowing for the advancement of precision medicine diagnostics and therapeutics.

The origins of precision medicine are not completely clear, in part due to the evolution of the term over time (Phillips 2020). However, one of the first fields to use a precision medicine approach to treat human disease was transfusion medicine, where the discovery of blood types in the early 1900s revolutionized blood transfusions, allowing for matching of donor and recipient blood types, and avoiding complications associated with mismatched donor and recipient blood (Dance 2016; Giangrande 2000; Hodson 2016). Since that time, precision medicine has evolved considerably to include novel approaches to prevention, diagnosis, intervention, and treatment, all of which are changing the landscape of medicine. Many fatal illnesses now have precision medicine therapies that are prolonging life and improving quality of life for patients, such as gene therapy for infants with spinal muscular atrophy (SMA) type I — a disease once uniformly fatal before age 2. Children with SMA type I treated with gene therapy are now living longer and having far fewer serious respiratory complications needing invasive respiratory support, which has been life changing for these patients and their families (Singh et al. 2017). Building on its success and promise for the future, precision medicine is now a field that has significant support from research and clinical funding agencies, government administrations, and within the general population, including private donors and politicians. The purpose of this minireview is to highlight how machine learning can be an integral tool for the future of precision medicine. This review will primarily focus on genetics and genomics, "big data", and state-of-the-art machine learning applications, and will also discuss ethical and legal considerations.

Genetics, genomics, and precision medicine

From the discovery of DNA in 1869 by Friedrich Miescher and the first descriptions of DNA by Watson, Crick, and Franklin in 1953, to the first discovery of specific genetic mutations for colour vision and cystic fibrosis in the late 1980s, to the incredible understanding we now have of the genetic basis of health and disease, the fields of genetics and genomics have advanced at an exponential pace. Genetics, the study of genes and their roles in inheritance, and genomics, the study of a person's genome and the interactions within the genome and the greater environment, have both played significant roles in the launch of the precision medicine revolution, and as a result much of the initial focus of precision medicine has largely involved genetics and genomics. Indeed, the majority of available precision medicine data comes from the fields of genetics and genomics (Grainger 2016). This has also been possible due to the significant decreases in costs and time to conduct genetic testing — the first genome was sequenced in 2001 and took over a decade and an estimated US$3 billion to complete, while today a genome can be sequenced within 24 h for approximately US$1000 (Hodson 2016). There is no question that genetics and genomics have fuelled the field of precision medicine and will continue to be integral to the future of the field.

The "omics" revolution

However, precision medicine is rapidly moving to include data from other fields in addition to genetics and genomics (Peck 2018). Over time, refinement of techniques has allowed inclusion of data from a variety of other omics sources, including epigenetics "epigenomics", protein "proteomics", metabolics "metabolomics", radiology "radiomics", pharmacology "pharmacomics", microbiome studies "microbiomics", "environmental omics", and others. For this reason, inclusion of data from multiple domains is sometimes referred to as "multi-omics". With the creation of these complex datasets with large volumes of information from multiple sources, new data science methods to process, understand, and utilize this information were needed as these complex and deep phenotyping datasets are often difficult or impossible to analyze without the help of data science and increasing computing power and technology. Through this, the field of machine learning emerged as one important tool. Machine learning is a form of artificial intelligence that involves training computer models to process and understand data and uses the identified complex patterns to make classifications or predictions on new cases. Machine learning has revolutionized many aspects of our daily life already and will also be an integral tool for the future of precision medicine.

A short introduction to machine learning

The rise of "intelligent" machines and technology, known as artificial intelligence, was once only fantasy described by philosophers, artists, and science fiction writers; however, artificial intelligence is now part of daily life and is a cornerstone of medicine and research (Bruce 2005). Machine learning is a branch of artificial intelligence where computer-based models are created to identify and learn patterns in high-dimensional data to create prediction and classification models based on the training data. The term machine learning was first popularized in the 1950s by Arthur Samuel working at

IBM. Since that time, machine learning has evolved considerably (John and Edward 1990). Machine learning can be further subdivided into supervised and unsupervised learning (Zou et al. 2019) as well as reinforcement learning. Reinforcement learning models are trained based on direct reward and punishment as feedback for positive and negative performance. A positive feedback (reward) essentially trains the machine learning model to repeat the decision in future, while a negative feedback (punishment) trains the machine learning model to avoid the decision made in future. Due to the direct feedback needed, reinforcement learning plays a rather small role for precision medicine approaches compared to supervised or unsupervised machine learning methods.

Unsupervised machine learning aims to uncover patterns in unlabeled data. In doing so, it can automatically identify clusters of similar cases within a dataset. Once different clusters are identified they can be visualized or further analyzed, for example, using standard statistics. Unsupervised machine learning methods can be particularly helpful to answer questions such as "Are there different types of a disease?" or "How much does a patient with a disease differ from normal subjects?". Popular unsupervised machine learning models include principal component analysis, k-nearest-neighbours, or variational autoencoders (an unsupervised deep learning architecture). Contrary to this, supervised machine learning techniques aim to identify patterns in multidimensional data based on labelled data (e.g., healthy vs. disease or outcome scores). More precisely, a training dataset with ground truth labels is typically used to build a model and optimize performance of the machine learning model for the desired outcome (Husi 2019; Zou et al. 2019). The uncovered (learnt) patterns can then be used to classify new datasets or make data-driven, patient-individual predictions. Machine learning classification models are used to classify datasets, while regression models are typically used to predict continuous outcome scores. Popular supervised statistical and machine learning techniques include, for example, support vector machine, random forests, linear models, and deep neural networks. In many cases, corresponding machine learning models are available for both problems (e.g., support vector machines for classification and support vector regression for continuous outcome scores).

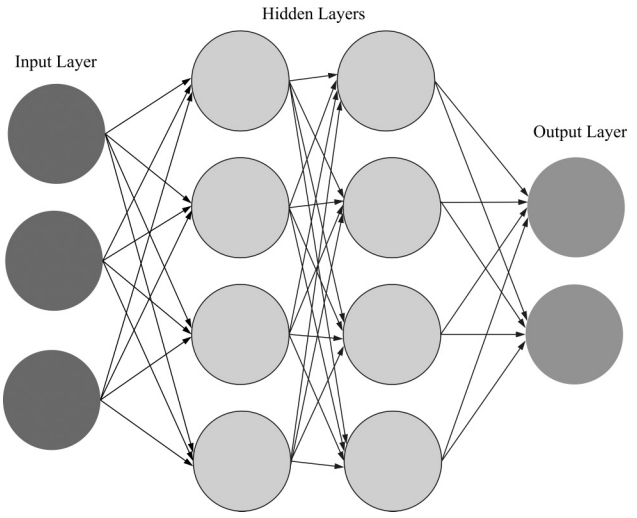
In order for machine learning techniques to be successfully applied, several aspects should be considered. First, the input data must be of high quality with small artifact or noise levels. Second, and even more important than a low noise level, is the correctness of the ground truth labels. While machine learning models can deal with noisy data to some extent, wrong labels can downgrade the performance of a machine learning model considerably and cannot be easily identified during training. Correctness of the ground truth labels must be ensured during the data curation process, such as by having correct diagnosis and diagnostic labels. Generally, errors in the ground truth labels have a more significant negative effect on the accuracy of machine learning models compared to other noise in the data. Third, like many statistical methods, most machine learning models also require a training set without missing features. Thus, it is important that the training sets are as complete as possible. While data augmentation methods can be used to fill in the missing data, which can range from random imputation to more advanced machine learning based algorithms, this does usually not lead to the same performance compared to using a complete dataset for training. Fourth, larger datasets are generally preferred, as this enables the machine learning model to learn the true variation in the data with a reduced risk that a few outliers affect the model negatively. However, collecting large enough datasets is one of the biggest hurdles to develop sophisticated machine learning models for precision medicine approaches, especially in the context of rare diseases. Fifth, in classification algorithms, precision and accuracy are both important, and, therefore, machine learning models should be optimized considering both aspects. Finally, every machine learning model should be validated prior to using it for computer-aided diagnosis support or clinical treatment decision making. For example, a machine learning model can be trained using datasets containing cases diagnosed by a physician. After this, new cases can be independently analyzed by a physician and the machine learning model so that the results can be compared to validate the model (Kononenko 2001).

Within supervised learning, there are multiple classification and regression models that can be used to solve a specific problem, all of which have strengths and weaknesses when considering a specific clinical problem. A detailed description of these methods, including specifics of each of the machine learning subtypes, is beyond the scope of this review and the interested reader is referred to textbooks such as *Applied Predictive Modelling* (Kuhn and Johnson 2013) or review papers including Angermueller et al. (2016), Ching et al. (2018), and a recently published tutorial for clinicians (Lo Vercio et al. 2020) for more details.

Machine learning using artificial neural networks

Briefly described, artificial neural networks are inspired by the connectivity of neurons and structures found within the human brain. Artificial neural networks consist of an input layer, multiple hidden layers, and an output layer (Fig. 1). Each layer consists of multiple artificial neurons that are typically connected to the neurons in the next layer (this is typically referred to as a feed-forward network) via so-called weights, which replicate the axons and dendrites that connect brain

Fig. 1. Visualization of a fully connected feed-forward artificial neural network with two hidden layers.



cells. The number of neurons in the input layer is typically equal to the number of input features while the number of neurons in the output layer is typically equal to the number of classes in case of a classification problem, for example patients with or without a disease. By increasing the number of hidden layers and neurons in each hidden layer, the network has increasing ability to solve more complex non-linear problems. However, the network also becomes more difficult to optimize and train (Husi 2019). Artificial neural networks with many hidden layers are typically referred to as deep neural networks.

The history of artificial neural networks dates back to 1943 when Walter Pitts and Warren McCulloch presented the first computer model of a neuron. Between 1965 and 1971, Alexey Ivakhnenko first described small artificial neural networks composed of up to eight layers with interconnected artificial neurons. However, training these artificial neural networks was computationally very expensive given the hardware and algorithms available at that time and as a result the popularity of this machine learning type was limited. The next breakthrough was achieved in the 1980s, when James Hopfield presented the first version of a recurrent neural network (Hopfield 1982, 1984) and Geoffrey Hinton and colleagues (Rumelhart et al. 1986) popularized backpropagation for training of neural networks, which enabled a considerably more efficient training of complex artificial neural networks. However, it was not until early 2000 that new hardware, especially specialised graphical processing units, and even more efficient algorithms became available that allowed to develop and train artificial neural networks with complex architectures and many hidden layers that soon after started to outperform many "classical" machine learning techniques such as support vector machines and random forests.

Many different specialized neural networks architectures have been developed, and recently deep learning machine learning models have been outperforming traditional machine learning models for many problems, even outperforming humans in many tasks. For example, recurrent neural networks, including long short-term memory deep learning, are utilized to process sequential time-series data, while convolutional neural networks are powerful tools to solve complex signal and image processing tasks such as the automatic image-based tumor staging. Convolutional neural networks apply convolutional filters in the first layers of the deep neural network that can automatically identify important temporal or spatial features in datasets such as whole genome data or three-dimensional imaging data. The convolutional filters are optimized in the same fashion as the network weights are optimized. Thus, one does not need to handcraft feature extraction filters, but instead the raw dataset is used directly as an input to the deep convolutional neural network and the optimal temporal or spatial features for the machine learning problem at hand are automatically identified.

Given their power and promise, deep learning methods such as convolutional neural networks will likely play an increasingly important role for precision medicine applications in future.

Advantages and limitations of machine learning

Machine learning, and in particular artificial neural networks, have several advantages for data analysis in the era of precision medicine. One major benefit of most machine learning models is that no strict assumptions about the data distribution (e.g., normal distribution) are made. Within this context, most machine learning models can easily combine multi-omics data, including binary, categorial, discrete, and continuous variables without the need of extensive data preprocessing. Due to the regularization used in many machine learning methods, most machine learning models can also handle noisy data and large variances within the dataset comparatively well, although of course less noise is always preferred. Another benefit of machine learning models is that there are specialized types and architectures that can be trained on small datasets, especially those for which the number of features considerably outnumber the number of observations. At the same time, complex machine learning models can identify multi-faceted, non-linear patterns in the training data, which might not be obvious to human observers or simple linear models.

Conversely, machine learning also has limitations. For example, the ability to identify complex, non-linear patterns also comes along with reduced interpretation capabilities compared to simple linear models. For this reason, machine learning models are often criticized as being a "black box". Within this context, it is important to ensure that such machine learning models do not

overfit the training data, which comes along with decreased generalizability. Overfitting in this case means that the machine learning model performs considerably better on the training data compared to the test data. In other words, the machine learning model adapts too well to the training data. Generally, the more complex a machine model is, the higher is the chance of overfitting. For this reason, deep neural networks with thousand to millions of parameters are especially prone to overfitting. Likewise, machine learning models can also underfit the training data, which is associated with reduced accuracy in the training as well as testing data (Zou et al. 2019). To ensure that the developed machine learning models do not over- or underfit the data, a completely independent test set should generally be used to evaluate the developed machine learning model.

Machine learning and genomics

Many machine learning methods have been successfully applied to a wide variety of genomics data, in particular due to large dataset sizes and complexity of data that are challenging to process with traditional statistical methods or classical linear machine learning models. The power of machine learning in genomics is best illustrated by highlighting a few unique examples where machine learning models applied to genomics data successfully overcame challenges posed by these large and complex datasets and resulted in clinically relevant outcomes. In the following, some selected examples will be described in further detail to showcase the broad applicability of machine learning across multiple disciplines.

For example, machine learning has been used extensively in genome-wide association studies (GWAS), where genomic information is examined to identify variants associated with a trait or phenotype. Wei et al. trained and evaluated a support vector machine model based on GWAS data to identify patients with Type I diabetes and concluded from their results that a genotype-based disease risk assessment may be possible for diseases for which single nucleotide polymorphism (SNP) arrays capture a large risk proportion (Wei et al. 2009). Romagnoni et al. compared multiple classical machine learning models and their ability to classify patients with Crohn's disease using genome-wide genotyping data (Romagnoni et al. 2019). Using 18 227 Crohn's disease patients and 34 050 healthy controls, they found that non-linear classification methods such as boosted trees classifiers or neural networks with more than one hidden layer perform considerably better compared to linear models such as a simple logistic regression.

Using data generated from previous patients treated for a disease, machine learning models can identify future patients who may benefit from a specific treatment. Machine learning models can also be applied to multi-modal data acquired through electronic medical records (EMRs) or other curated data sources to identify patients with conditions that may benefit from early treatment or participation in randomized control trials of novel interventions (Rajkomar et al. 2019). For example, Dong et al. developed and evaluated a support vector machine model for anticancer drug sensitivity prediction using genomic data, and demonstrated that response to cancer treatment could be predicted based on genomics, which if clinically applied could help avoid unnecessary treatments in non-responders in favour of the most effective treatment based on the patient's genome (Dong et al. 2015). Additionally, machine learning methods, including deep learning methods, have been used in pharmacogenomics, which is a relatively new research field, which may allow for mechanistic prediction of drug response and may help inform personalized drug design (Kalinin et al. 2018). A detailed description of these methods is outside the scope of this minireview; interested readers are, for example, directed to the excellent review by Kalinin et al. (2018).

Another application of machine learning is to identify novel biomarkers for specific diseases. Such biomarkers can support early disease detection, prediction of treatment response, and prognostication of disease outcome. Many of these studies have been aided by the large amounts of publicly available databases such as the UK Biobank or the Alzheimer's Disease Neuroimaging Initiative. For example, one application in this context used "big data" from over 11 000 tumours from 33 different types of cancer in combination with unsupervised machine learning in an attempt to understand their similarities and differences (Hoadley et al. 2018; Weinstein et al. 2013). This ongoing study has improved our understanding of the way cancer mutates from origin cells and the factors that modulate tumours, including immunologic factors, with the ultimate goal of identifying therapeutic targets for cancer treatment (Hoadley et al. 2018). Other models are being developed for disease prognostication, such as utilizing genetic data in a machine learning model to identify biomarkers linked with better survival rates in patients with bladder cancer, thereby allowing for classification of patients into survival subtypes (Poirion et al. 2018). Likewise, Szymczak et al. developed a new variable selection method for GWAS data using a random forest model (a machine learning model related to boosted trees), which ranks SNPs according to their predictive power and can be applied to a variety of GWAS datasets (Szymczak et al. 2016). Due to fact that random forest models are known to be very powerful when used for datasets with a large number of variables and considerably fewer observations, this machine learning model has been applied to various other genomic data analysis tasks in addition to classification and feature selection such as pathway analysis, genetic association, and epistasis detection (Chen and Ishwaran 2012).

Progress has also been made in understanding other heritable conditions beyond cancer using novel deep learning models. For example, Montaez et al. developed a deep learning model to classify patients with obesity based on statistically significant SNPs associated with obesity phenotype and showed that this model can successfully identify relevant SNPs and interactions between them for this specific classification task (Montaez et al. 2018). A deep convolutional neural network model called DeepVariant was developed and applied to next-generation sequencing data to help identify genetic variants within a whole genome, including the difficult task of determining true variants from sequencing errors, which outperformed existing techniques (Poplin et al. 2018). Another deep learning algorithm, DeepSEA, was specifically developed to help identify functional effects of non-coding variants, which is where the majority of disease-associated SNPs are found within the genome. Determining the functional effects of specific non-coding variants is a significant challenge due to the large number of such variants within the genome, most of which are poorly understood (Zhou and Troyanskaya 2015). Using genomic sequences as the input data, the authors trained DeepSEA using chromatin profiles, which play an essential role in the regulation of gene transcription and epigenetics (Minard et al. 2009), allowing to predict which SNPs result in a functional difference, with implications for human disease (Zhou and Troyanskaya 2015). As another example, Yin et al. developed a convolutional neural network that they applied to genetic data from patients with amyotrophic lateral sclerosis (ALS), for which the genetic profile is complex and poorly understood. Using their convolutional neural network, they were able to successfully identify patients at risk of ALS based on identification of ALS-associated promoter regions (B. Yin et al. 2019).

However, many previous studies have also found that heritability only accounts for a small proportion of common diseases and have revealed both the limitations of GWAS and the complexity of genetic inheritance, including the need to consider epigenetic mechanisms (Nicholls et al. 2020). Novel machine learning models have recently been developed specifically to help identify and understand epigenetics and epigenomics, the study of phenomena that modify gene expression without modifying the underlying DNA sequence, as epigenetics and epigenomics add another layer of complexity to already complex genetics and genomics datasets. For example, DeepBind, a convolutional neural network, was designed to predict the sequences recognized by DNA and RNA binding proteins, which play an essential role in gene regulation (Alipanahi et al. 2015). Identifying these sequences allows for further development of biologic models and identification of genetic variants implicated in human disease. Similarly, DeepChrome and DeepHistone were developed to understand if histone protein modifications, the building blocks of chromatin that can alter chromatin itself, have combinational effects on gene expression, and act as markers for epigenetic changes that may modify health and disease (Singh et al. 2016; Q. Yin et al. 2019).

Integration of multi-modal data using machine learning

One major aim of using machine learning combining multiple omics data sources is to improve our understanding of the genotype–phenotype relation. One modality that is of special interest in this context is medical imaging, which is the largest data source by far in the health care system today. For this reason, it is also very challenging to identify relationships between deep genetic and high-dimensional imaging features and more precisely how genetic variations alter the anatomy or function of organs, which can help to improve understanding of disease phenotypes. This is the primary aim of a very active research field referred to as imaging genetics. To provide some examples for context, Jonsson et al. developed a convolutional neural network to estimate an individual's brain age from high-resolution T1-weighted magnetic resonance imaging (MRI) datasets of the brain (Jonsson et al. 2019). In the next step, they calculated the difference between the chronological and predicted brain age and investigated associations of this difference with GWAS data. Applying this approach to data from the UK Biobank, the authors identified two sequence variants (rs1452628-T and rs2435204-G) that are associated with atypical brain aging (larger difference between chronological and predicted brain age). Another interesting method combining genetic and imaging data was recently presented by Hallgrímsson et al., who developed and evaluated machine learning models for automatic syndrome diagnosis using 3D facial images of individuals with craniofacial syndromes. Their method achieved a balanced accuracy of 73% in identifying genetically confirmed craniofacial syndromes in a database of more than 7000 subjects (Hallgrímsson et al. 2020). Many worldwide initiatives are ongoing, such as the IMAGEN study, which is investigating biological, psychological, and environmental factors that influence brain development and mental health, using brain imaging and genetics (Mascarell Maričić et al. 2020). This area presents significant promise for the future of precision medicine.

In addition to using machine learning to investigate the genotype–phenotype relationship, machine learning can also be used to combine multi-omics data for disease diagnosis, prognosis, and treatment. Such a multi-parametric approach merging multi-omics data using advanced machine learning methods can incorporate a broad spectrum of important mechanistic information across disciplines in medicine, for example,

combining genetics and metabolomics (Zampieri et al. 2019). For example, Yang et al. showed that combining functional MRI and SNP data can considerably increase the accuracy of a support vector machine model to classify patients with schizophrenia compared to using the imaging or SNP data alone for this purpose (Yang et al. 2010). Khanna et al. developed a boosted tree machine learning model to predict the time until Alzheimer's disease diagnosis in normal subjects and patients with mild cognitive impairment based on genotype information, neuroimaging, and clinical data, including neuropsychological measures, acquired by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Khanna et al. 2018). Additionally, Bayesian networks were used to identify associations between neuro-psychological assessment scores, single genetic variants, pathways, and imaging features in the data. Machine learning has also been utilized for treatment response prediction. For example, Lin et al. developed a deep learning model to predict antidepressant treatment response in patients with major depressive disorder based on SNP, demographic, and clinical data, achieving a sensitivity of 75% and specificity of 69% (Lin et al. 2018).

Despite machine learning being optimally suited to make use of deep multi-omics data, many machine learning studies continue to be limited to a single data modality, potentially due to lack of collaboration between specialists and the associated problems collecting true multi-omics data across disciplines. Thus, there is still tremendous potential for machine learning models for precision medicine to make full use of multi-omics data in research and patient care.

Machine learning and the future of medicine

The practice of medicine has become increasingly complex, due in part to the discovery of relatively new fields such as genetics, metabolics, and immunology, the expansion of knowledge in others, the advent of new technologies ranging from imaging modalities and interventional techniques to genetic testing, and the pace of new therapies becoming available. Physicians are becoming more and more subspecialized to stay on top of new developments in their already specialized area of practice. However, despite the technological advances postulated to begin replacing physicians and their "intellectual functions" 50 years ago (Schwartz 1970), the clinician is still necessary and valuable for the practice of medicine. In particular, the interpretation and explanation of complex results and compassionate understanding of the patient in the context of a patient's life are all essential tasks that are beyond the scope of any machine learning algorithm (Eyal et al. 2019). For all its promises, precision medicine will not replace the clinicians any time soon but can support and help clinicians considerably with diagnosis, prognosis, and treatment using patient-individual data analyzed using machine learning techniques.

Ethical, legal, and moral considerations of precision medicine

There are several important ethical, legal, and moral considerations of machine learning in the context of precision medicine that must be considered. The first is data privacy. Machine learning models typically perform better with more datasets being used for model training. However, to collect a training set sufficiently large enough for machine learning model training generally requires that the datasets are collected at multiple locations and transferred to a central location for storage and processing. Therefore, it is essential that the datasets shared do not include any patient-identifying information to ensure patient privacy. In future, transfer of data and the associated privacy, ethical, and legal considerations could be overcome using novel distributed machine learning approaches. This would involve a travelling machine learning model, which trains locally on the data available, before moving to the next center with available data (Tuladhar et al. 2020). In this way, data does not have to leave the contributing center, and the machine learning model would not contain any patient-identifying information before moving to another center. However, more research is needed to produce machine learning models that are as effective and accurate as those that are trained using centralized data.

Another major concern is that the decisions of many machine learning models are not easy to understand and interpret as noted above, especially with increased complexity of the input data and within the machine learning models themselves. This raises legal and ethical concerns, as research and physicians and other clinicians may not fully understand how a machine learning model came to a specific conclusion about a patient's care. Current research in the machine learning domain is focussed on increasing the interpretability and thereby decreasing the "black box" nature of machine learning models, which will play a crucial role in the acceptance of such models within clinical decision making, as it will allow the physician or clinician to understand how the machine learning model came to a specific decision that then modifies care of the patient. An interesting consideration is the responsibility for error induced by the machine learning model, as physicians can be held legally responsible for their decisions, but it is unclear who can or should be held responsible when a machine learning model makes the wrong decision. Agencies approving medical devices and medical software tools, such as the Food and Drug Administration in the United States, are evaluating the safety and efficacy of an increasing number of medical devices that include artificial intelligence. Currently, all medical devices that employ machine learning must be marketed as a diagnosis support tool and not as tools for an autonomous diagnosis making, as their aim is to

support physicians and other clinicians in the clinical decision-making process.

Additionally, there are some technical limitations to machine learning that can raise general ethical, legal, and moral issues. Generally, machine learning models are only as good as the datasets they were trained with. However, datasets used for training are often incomplete, noisy, and may have inherent biases. This also means that the models that are produced using such datasets might be not generalizable when applied to datasets from other health care centers or geographic locations, rendering the machine learning model useless for a broad application. The easiest way to solve this problem is to increase the training data, which is often limited by the legal and logistical issues described above.

Another major concern with precision medicine is the financial cost. While becoming more affordable as technology advances, genetic testing, imaging, and other clinical investigations are still costly to conduct. Furthermore, precision medicine therapies are amongst the most expensive, for example, the cost of the life prolonging gene therapies for SMA type I (described in the Introduction) is in the range of millions of dollars. In publicly funded medical systems, these costs must be carefully considered as resources are limited. Privatized for-profit medical systems favour the economically advantaged, as individuals who can afford to pay high costs receive in-depth workups and access to testing, treatments, and interventions that are out of reach for many. This leads to significant disparities in access to care based on socioeconomic status that are unacceptable. Additionally, these in-depth workups can have unintended consequences, such as incidental findings that may lead to further testing, which may be invasive and unnecessary, as well as induce patient stress in the face of diagnostic uncertainty. Within this context, it is essential that future machine learning models can identify the most important features for a specific clinical question so that unnecessary assessments can be prevented or at least reduced, thereby maximizing benefit and minimizing harm to patients. Through this, the cost of obtaining the data could be offset in reductions in side effects of unhelpful treatments, misdiagnosis, and improved patient management.

Another major concern is the bias that may exist within precision medicine datasets, in particular with respect to genetic and genomic data, for example, due to patient selection. These biases may not only negatively affect the accuracy and generalizability of trained machine learning models but also aggravate economic and other disparities, and contribute to systemic racism, ableism, sexism, and classism (Ferryman and Pitcan 2018). Additionally, the financial and practical aspects of implementing precision medicine limit them primarily to the developed world. As a consequence, those in the developing and third world are often excluded from its analyses and benefits (Mentis et al. 2018). A concerted effort to include such populations must be made to ensure global equity and further our understanding and treatment of health and disease for all of humanity.

An interesting question is whether or not discoveries made by precision medicine will have individual impacts on human behaviour with respect to disease prevention. If a patient knows they are at risk for heart disease or obesity based on their underlying genetics or due to epigenetic factors, will they change their behaviour? Additionally, such discoveries may create "patients-in-waiting" who are not yet sick but have been identified as high risk, which may cause high levels of confusion and stress for these individuals and their families (Eyal et al. 2019). To this end, those applying precision medicine clinically must never lose sight of the individual in front of them and the context within which they live their lives.

Conclusion

Precision medicine is a novel approach to research and clinical practice that utilizes data from multiple sources to understand and treat human disease. Machine learning methodologies, including the emerging deep learning models, are an important component of the analysis and processing of the multi-omics data that allow for the classification and prediction of outcomes for individuals and populations. Genetics and genomics have been essential to the evolution of precision medicine and are well-suited to machine learning technology, and these data in combination with other assessments will be essential to the continued evolution of these fields. Precision medicine and machine learning have already contributed to significant advances in our understanding of human health and disease and hold great potential for the future of all humanity.

Funding

This work was supported by the Canada Research Chairs program and the River Fund at Calgary Foundation.

Conflicts of interest statement

Authors declare no conflict of interest.

References

Alipanahi, B., Delong, A., Weirauch, M.T., and Frey, B.J. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol. **33**(8): 831–838. doi:10.1038/nbt.3300. PMID:26213851.

Angermueller, C., Pärnamaa, T., Parts, L., and Stegle, O. 2016. Deep learning for computational biology. Mol. Syst. Biol. **12**(7): 878. doi:10.15252/msb.20156651. PMID:27474269.

Bruce, G.B. 2005. A (Very) Brief History of Artificial Intelligence. AI Magazine, **26**: 4. doi:10.1609/aimag.v26i4.1848.

Chen, X., and Ishwaran, H. 2012. Random forests for genomic data analysis. Genomics, **99**(6): 323–329. doi:10.1016/j.ygeno.2012.04.003. PMID:22546560.

Ching, T., Himmelstein, D.S., Beaulieu-Jones, B.K., Kalinin, A.A., Do, B.T., Way, G.P., et al. 2018. Opportunities and obstacles for deep learning in biology and medicine. J. R. Soc. Interface, **15**(141): 20170387. doi:10.1098/rsif.2017.0387. PMID: 29618526.

Dance, A. 2016. Medical histories. Nature, **537**(7619): S52–53. doi:10.1038/537S52a.

Dong, Z., Zhang, N., Li, C., Wang, H., Fang, Y., Wang, J., and Zheng, X. 2015. Anticancer drug sensitivity prediction in cell lines from baseline gene expression through recursive feature selection. BMC Cancer. **15**(1): 489. doi:10.1186/s12885-015-1492-6. PMID:26121976.

Eyal, G., Sabatello, M., Tabb, K., Adams, R., Jones, M., Lichtenberg, F.R., et al. 2019. The physician-patient relationship in the age of precision medicine. Genet. Med. **21**(4): 813–815. doi:10.1038/s41436-018-0286-z. PMID:30214065.

Ferryman, K., and Pitcan, M. 2018. Fairness in precision medicine. Data Soc.

Giangrande, P.L.F. 2000. The history of blood transfusion. Brit. J. Haematol. **110**(4): 758–767. doi:10.1046/j.1365-2141.2000.02139.x. PMID:11054057.

Grainger, D. 2016. The multi-omics revolution. J. Pers. Med. 81–86.

Hallgrímsson, B., Aponte, J.D., Katz, D.C., Bannister, J.J., Riccardi, S.L., Mahasuwan, N., et al. 2020. Automated syndrome diagnosis by three-dimensional facial imaging. Genetics in Medicine. **22**: 1682–1693. doi:10.1038/s41436-020-0845-y. PMID:32475986.

Hoadley, K.A., Yau, C., Hinoue, T., Wolf, D.M., Lazar, A.J., Drill, E., et al. 2018. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell, **173**(2): 291–304.E6. doi:10.1016/j.cell.2018.03.022. PMID:29625048.

Hodson, R. 2016. Precision medicine. Nature, **537**(7619): S49–S49. doi:10.1038/537S49a. PMID:29733714.

Hopfield, J.J. 1982. Neural networks and physical systems with emergent collective computational abilities. Proc. Natl. Acad. Sci. U.S.A. **79**(8): 2554–2558. doi:10.1073/pnas.79.8.2554. PMID:6953413.

Hopfield, J.J. 1984. Neurons with graded response have collective computational properties like those of two-state neurons. Proc. Natl. Acad. Sci. U.S.A. **81**(10): 3088–3092. doi:10.1073/pnas.81.10.3088. PMID:6587342.

Husi, H. 2019. Computational Biology. Codon Publications. Brisbane (AU).

John, M., and Edward, A.F. 1990. In Memoriam: Arthur Samuel: Pioneer in machine learning. AI Magazine. **11**(3): 10–11. doi:10.1609/aimag.v11i3.840.

Jonsson, B.A., Bjornsdottir, G., Thorgeirsson, T.E., Ellingsen, L.M., Walters, G.B., Gudbjartsson, D.F., et al. 2019. Brain age prediction using deep learning uncovers associated sequence variants. Nat. Commun. **10**(1): 5409. doi:10.1038/s41467-019-13163-9.

Kalinin, A.A., Higgins, G.A., Reamaroon, N., Soroushmehr, S., Allyn-Feuer, A., Dinov, I.D., et al. 2018. Deep learning in pharmacogenomics: from gene regulation to patient stratification. Pharmacogenomics, **19**(7): 629–650. doi:10.2217/pgs-2018-0008. PMID:29697304.

Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M.A., Hofmann-Apitius, M., and Fröhlich, H. 2018. Using multi-scale genetic, neuroimaging and clinical data for predicting alzheimer's disease and reconstruction of relevant biological mechanisms. Sci. Rep. **8**(1): 11173. doi:10.1038/s41598-018-29433-3. PMID:30042519.

Kononenko, I. 2001. Machine learning for medical diagnosis: history, state of the art and perspective. Artif. Intell. Med. **23**(1): 89–109. doi:10.1016/S0933-3657(01)00077-X. PMID:11470218.

Kuhn, M. and Johnson, K. 2013. Applied predictive modeling. Springer. Vol. 26.

Lin, E., Kuo, P.-H., Liu, Y.-L., Yu, Y.W.-Y., Yang, A.C., and Tsai, S.-J. 2018. A deep learning approach for predicting antidepressant response in major depression using clinical and genetic biomarkers. Front. Psychiatry **9**: 290. doi:10.3389/fpsyt.2018.00290. PMID:30034349.

Lo Vercio, L., Amador, K., Bannister, J., Crites, S., Gutierrez, A., MacDonald, M., and Forkert, N., 2020. Supervised machine learning tools: a tutorial for clinicians. J. Neural. Eng. **17**(6): 062001. doi:10.1088/1741-2552/abbff2. PMID:33036008.

Mascarell Maričić, L., Walter, H., Rosenthal, A., Ripke, S., Quinlan, E.B., Banaschewski, T., et al. 2020. The IMAGEN study: a decade of imaging genetics in adolescents. Mol. Psychiatry, **25**: 2648–2671. doi:10.1038/s41380-020-0822-5. PMID:32601453.

Mentis, A.-F.A., Pantelidi, K., Dardiotis, E., Hadjigeorgiou, G.M., and Petinaki, E. 2018. Precision medicine and global health: the good, the bad, and the ugly. Front. Med. **5**(67). doi:10.3389/fmed.2018.00067. PMID:29594124.

Minard, M.E., Jain, A.K., and Barton, M.C. 2009. Analysis of epigenetic alterations to chromatin during development. Genesis (New York, N.Y), **47**(8): 559–572. doi:10.1002/dvg.20534. PMID:19603511.

Montaez, C.A.C., Fergus, P., Montaez, A.C., Hussain, A., Al-Jumeily, D., and Chalmers, C. 2018. Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs. Paper presented at the 2018 International Joint Conference on Neural Networks (IJCNN). 8-13 July 2018. doi:10.1109/IJCNN.2018.8489048.

Nicholls, H.L., John, C.R., Watson, D.S., Munroe, P.B., Barnes, M.R., and Cabrera, C.P. 2020. Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. Front. Genet. **11**: 350–350. doi:10.3389/fgene.2020.00350. PMID:32351543.

Peck, R.W. 2018. Precision medicine is not just genomics: the right dose for every patient. Annu. Rev. Pharmacol Toxicol. **58**: 105–122. doi:10.1146/annurev-pharmtox-010617-052446. PMID:28961067.

Phillips, C.J. 2020. Precision medicine and its imprecise history. Harvard Data Science Review. **2**(1). doi:10.1162/99608f92.3e85b56a. PMID:31932784.

Poirion, O.B., Chaudhary, K., and Garmire, L.X. 2018. Deep Learning data integration for better risk stratification models of bladder cancer. AMIA Jt. Summits Transl, Sci, Proc. **2017**: 197–206. PMID:29888072.

Poplin, R., Chang, P.-C., Alexander, D., Schwartz, S., Colthurst, T., Ku, A., et al. 2018. A universal SNP and small-indel variant caller using deep neural networks. Nat. Biotechnol. **36**(10): 983–987. doi:10.1038/nbt.4235. PMID:30247488.

Rajkomar, A., Dean, J., and Kohane, I. 2019. Machine learning in medicine. N. Eng. J. Med. **380**(14): 1347–1358. doi:10.1056/NEJMra1814259. PMID:30943338.

Romagnoni, A., Jégou, S., Van Steen, K., Wainrib, G., Hugot, J.-P., and Peyrin-Biroulet, L. International Inflammatory Bowel Disease Genetics Consortium. 2019. Comparative performances of machine learning methods for classifying Crohn Disease patients using genome-wide genotyping data. Sci. Rep. **9**(1): 10351. doi:10.1038/s41598-019-46649-z. PMID:31316157.

Rumelhart, D.E., Hinton, G.E., and Williams, R.J. 1986. Learning representations by back-propagating errors. Nature, **323**(6088): 533–536. doi:10.1038/323533a0.

Schwartz, W.B. 1970. Medicine and the computer. The promise and problems of change. N Engl. J. Med. **283**(23): 1257–1264. doi:10.1056/NEJM197012032832305. PMID:4920342.

Singh, N.N., Howell, M.D., Androphy, E.J., and Singh, R.N. 2017. How the discovery of ISS-N1 led to the first medical therapy for spinal muscular atrophy. Gene Ther. **24**(9): 520–526. doi:10.1038/gt.2017.34. PMID:28485722.

Singh, R., Lanchantin, J., Robins, G., and Qi, Y. 2016. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics, **32**(17): i639–i648. doi:10.1093/bioinformatics/btw427. PMID:27587684.

Szymczak, S., Holzinger, E., Dasgupta, A., Malley, J.D., Molloy, A.M., Mills, J.L., et al. 2016. r2VIM: A new variable selection method for random forests in genome-wide association studies. BioData Min. **9**(1): 7. doi:10.1186/s13040-016-0087-3. PMID:26839594.

Tuladhar, A., Gill, S., Ismail, Z., and Forkert, N.D. 2020. Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling. J. Biomed. Inform. **106**: 103424. doi:10.1016/j.jbi.2020.103424. PMID:32335226.

Wei, Z., Wang, K., Qu, H.-Q., Zhang, H., Bradfield, J., Kim, C., et al. 2009. From disease association to risk assessment: an optimistic view from genome-wide association studies on type 1 diabetes. PLoS Genet. **5**(10): e1000678. doi:10.1371/journal.pgen.1000678. PMID:19816555.

Weinstein, J.N., Collisson, E.A., Mills, G.B., Shaw, K.R.M., Ozenberger, B.A., Ellrott, K., et al. 2013. The cancer genome atlas pan-cancer analysis project. Nat. Genet. **45**(10): 1113–1120. doi:10.1038/ng.2764.

Yang, H., Liu, J., Sui, J., Pearlson, G., and Calhoun, V.D. 2010. A hybrid machine learning method for fusing fMRI and genetic data: combining both improves classification of schizophrenia. Front. Hum. Neurosci. **4**: 192–192. doi:10.3389/fnhum.2010.00192. PMID:21119772.

Yin, B., Balvert, M., van der Spek, R.A.A., Dutilh, B.E., Bohté, S., Veldink, J., and Schönhuth, A. 2019. Using the structure of genome data in the design of deep neural networks for predicting amyotrophic lateral sclerosis from genotype. Bioinformatics, **35**(14): i538–i547. doi:10.1093/bioinformatics/btz369. PMID:31510706.

Yin, Q., Wu, M., Liu, Q., Lv, H., and Jiang, R. 2019. DeepHistone: a deep learning approach to predicting histone modifications. BMC Genomics, **20**(2): 193. doi:10.1186/s12864-019-5489-4. PMID:30967126.

Zampieri, G., Vijayakumar, S., Yaneske, E., and Angione, C. 2019. Machine and deep learning meet genome-scale metabolic modeling. PLoS Comput. Biol. **15**(7): e1007084. doi:10.1371/journal.pcbi.1007084. PMID:31295267.

Zhou, J., and Troyanskaya, O.G. 2015. Predicting effects of noncoding variants with deep learning–based sequence model. Nat. Methods. **12**(10): 931–934. doi:10.1038/nmeth.3547. PMID:26301843.

Zou, J., Huss, M., Abid, A., Mohammadi, P., Torkamani, A., and Telenti, A. 2019. A primer on deep learning in genomics. Nat. Genet. **51**(1): 12–18. doi:10.1038/s41588-018-0295-5. PMID:30478442.