Molecular Diversity
https://doi.org/10.1007/s11030-021-10288-2

ORIGINAL ARTICLE

Prediction of antischistosomal small molecules using machine learning in the era of big data

Samuel K. Kwofie[1,2] · Kwasi Agyenkwa-Mawuli[1,2] · Emmanuel Broni[1,6] · Whelton A. Miller III[3,4,5] · Michael D. Wilson[3,6]

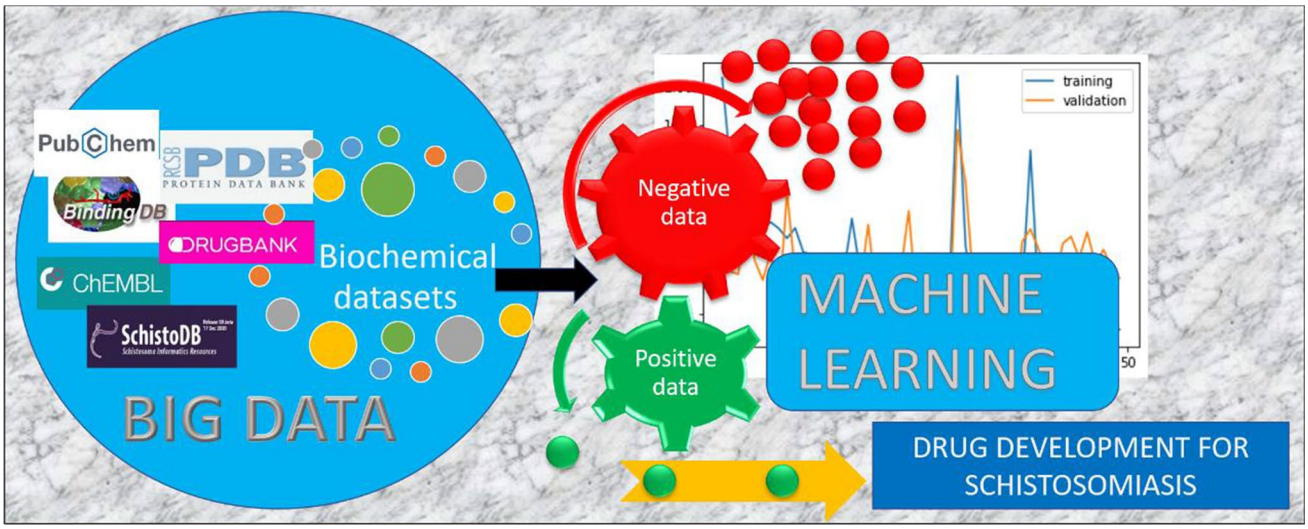
Received: 24 March 2021 / Accepted: 24 July 2021

Abstract

Schistosomiasis is a neglected tropical disease caused by helminths of the Schistosoma genus. Despite its high morbidity and socio-economic burden, therapeutics are just a handful with praziquantel being the main drug. Praziquantel is an old drug registered for human use in 1982 and has since been administered en masse for chemotherapy, risking the development of resistance, thus the need for new drugs with different mechanisms of action. This review examines the use of machine learning (ML) in this era of big data to aid in the prediction of novel antischistosomal molecules. It first discusses the challenges of drug discovery in schistosomiasis. Explanations are then offered for big data, its characteristics and then, some open databases where large biochemical data on schistosomiasis can be obtained for ML model development are examined. The concepts of artificial intelligence, ML, and deep learning and their drug applications are explored in schistosomiasis. The use of binary classification in predicting antischistosomal compounds and some algorithms that have been applied including random forest and naive Bayesian are discussed. For this review, some deep learning algorithms (deep neural networks) are proposed as novel algorithms for predicting antischistosomal molecules via binary classification. Databases specifically designed for housing bioactivity data on antischistosomal molecules enriched with functional genomic datasets and ontologies are thus urgently needed for developing predictive ML models.

Graphic abstract

This shows the application of machine learning techniques for the discovery of novel antischistosomal small molecules via binary classification in the era of big data.



Extended author information available on the last page of the article

Published online: 05 August 2021

Keywords Machine learning · Deep learning · Artificial intelligence · Big data · Drug discovery · Schistosomiasis · Classifiers · Binary classification

Background

Schistosomiasis is a fresh water-borne disease that affects almost 240 million people worldwide with more than 700 million people living in endemic areas. It is caused by infections with helminths belonging to the genus *Schistosoma*. The medically relevant species are *Schistosoma haematobium*, *S. mansoni, S. japonicum,* and *S. mekongi* [1]. The infection is prevalent in tropical and sub-tropical areas, in poor communities without potable water and adequate sanitation. The most severe clinical outcomes of schistosomiasis are female and male genital lesions caused by *S. haematobium*, which are associated with pain, infertility, and an increased risk of contracting HIV. These urogenital manifestations often do not resolve after clearance of infection. Other serious clinical manifestations of intestinal schistosomiasis include chronic blood loss, cognitive impairment, and diminished academic performance. It is estimated that schistosomiasis accounts for the loss of 1.4 million disability-adjusted life years (DALYs) [2] although this is likely to be an underestimate [3]. These debilitating symptoms that are linked with schistosomiasis, further exacerbate the already precarious socio-economic situations of affected populations [2, 4].

The main treatment of schistosomiasis is chemotherapy with praziquantel, which is safe and effective, but it is an old drug and the only one that is recommended for mass treatment of populations. Moreover, there have been reports of drug-resistant *Schistosoma* species [5, 6]. Other repurposed drugs such as oxamniquine, metrifonate, and artemether have also been used with limited success [7, 8]. This coupled with the lack of a licensed vaccine creates the need for continuous intensified efforts for drug discovery in schistosomiasis [9, 10].

Drug discovery in schistosomiasis follows the general pipeline which often takes years and has several challenges [11]. Hence, the number of antischistosomal drug candidates that reach clinical trials is unsatisfactory. Moreover, the financial return is low for the pharmaceutical industry despite the huge investments they make for developing new drugs [6]. These limitations lead to underinvestment in new antischistosomal drugs with the industry preferring the cheaper alternative of repurposing candidates from other infectious diseases [12]. With the increasing significance of drug discovery in schistosomiasis in the scientific community, the demand for novel approaches is also on the rise [6].

Challenges and limited funding available for neglected disease drug discovery and development have necessitated exploring less costly approaches. Integrating machine learning (ML) into drug development pipelines could be the approach that decreases drug development costs and time. For instance, ML can be used to make and test hypotheses before undertaking costly drug experiments for neglected diseases such as Schistosomiasis [13]. As a predictive computational technique, ML can be readily implemented early in the process as a means to reduce the number of compounds for screening which saves time and money for later stages [13].

The use of ML techniques in addressing the medical challenges of schistosomiasis and other neglected tropical diseases (NTDs) is nothing new [14]. AI training models have been used in the predictive modeling of the disease based on vector density [15]. ML approaches were used to predict persistent hotspots [16], the prognosis of patients with advanced schistosomiasis [17], and classify snail and schistosome images for disease control [17, 18]. Similarly, ML was used to assess the prevalence and transmission risk [10, 19]. Trained ML models were used to predict protein essentiality characteristics of *Schistosoma mansoni* [20]. Furthermore, ML models have been used to classify G protein-coupled receptors (GPCRs) of *S. haematobium* and *S. mansoni* [21].

For drug discovery purposes, the activity of statin against schistosomula with an ML-based algorithm was the basis for a new antischistosomal drug [22]. Also, cost-sensitive supervised learning models (J48, NB, and RF) were used to evaluate the biological activity of molecules against *S. mansoni* [23]. Therefore, this review addresses the need to utilize the data-driven approach of binary classification with machine learning to aid in the discovery of antischistosomal small molecules.

Big data

The use of big data and machine learning approaches to aid in the discovery of potentially novel bioactive molecules of therapeutic relevance for schistosomiasis is gaining prominence. The exponential growth of massive bioactive datasets and advanced artificial intelligence (AI) techniques, specifically deep learning (DL), provides an alternative path for drug development and optimization for schistosomiasis [24]. Big data have varied definitions [25] depending on the discipline. In this review, the aspect of big data under focus is biological and chemical data related to drug discovery [26]. Big data were defined as datasets that satisfied $\text{Log}(n \times p) \geq 7$, where n represented the number of statistical individuals and p was the number of variables for all papers

describing a dataset [27]. In general, big data refers to a collection of datasets so large and complex that they are difficult to process with traditional data analysis approaches [26]. The distinguishing features of big data have been dubbed the "Vs', some of which are huge *volume*, high *velocity*, high *variety*, and low *veracity* [25]. The number of these descriptive "Vs" vary from five [25] and up to ten [28]. These features describe the overall complexity, scale, growth, quality, value, and sheer expanse of big data.

Advancements in chemical synthesis and biological screening technologies in drug discovery have enabled the generation of large datasets [29]. An example of such an approach is high-throughput screening (HTS), which has contributed immensely to the growth of big data over the past decade for drug discovery. The combination of experimental HTS and robotic technologies has enabled the screening of thousands to millions of compounds and small molecules [26]. This development has pushed drug discovery into the big data era, hence the need for new robust computational techniques to fully exploit them.

Data sources

Most of the data generated by the aforementioned approaches are stored in various databases mostly online via data-sharing projects. Databases perused for schistosomiasis-related data include PubChem [30], ChEMBL [31], DrugBank [32], BindingDB [33], Protein Data Bank (PDB) [34], and helminth-specific databases such as SchistoDB [35]. There are several others, but these are the most popular freely available databases.

PubChem (https://pubchem.ncbi.nlm.nih.gov/) is an online repository for chemical structures and their biological activities. As of August 2020, it contained over 293 million depositor-provided substance descriptions, 111 million unique chemical structures, and 271 million bioactivity data points from 1.2 million biological assay experiments [36]. To obtain schistosomiasis-related data in PubChem, a search conducted on Schistosoma yielded 8 substances, 2 genes, 23 proteins, 125 pathways, 674 bioassays, and 22 871 related literature. A typical bioassay, AID 485,364 had over 350 000 compounds screened against *Schistosoma mansoni* TGR, which was used in a chemoinformatic study [23].

A similar resource ChEMBL (https://www.ebi.ac.uk/chembl/) is a manually curated database of bioactive molecules with drug-like properties [37]. Version 26 of ChEMBL contains 15 996 368 activity data, 13 377 targets, 1 221 311 assays, and 2 425 876 compound records. A search conducted with the "Schistosoma" query string yielded 10 targets and 513 assays from 39 published literature. The quantitative high-throughput screening (qHTS) functional assay CHEMBL1614161 for example had 28 392 compounds (28 341 small molecules and 51 unknown) screened against *Schistosoma mansoni.*

DrugBank (https://go.drugbank.com/) is a database for drugs and drug candidates along with their relevant mechanisms, interactions, and targets [32]. Version 5.1.8 contains 14 315 drugs, 4 885 unique targets, and 18 866 drug-target associations. A search on Schistosoma yielded 2 drugs (oxamniquine and praziquantel) and 1 target (schistosome calcium ion channels). This corroborates the dearth of datasets on therapeutics for schistosomiasis.

BindingDB (https://www.bindingdb.org/bind/index.jsp) is a web-based database for measured binding affinities and interactions of proteins considered targets and small drug-like molecules [33]. As of December 2020, it contained 2 096 653 binding data for 8 185 protein targets and 920 703 small molecules. It contained 7 assays on *Schistosoma* with all bearing ChEMBL IDs (Table 1).

Protein Data Bank (https://www.rcsb.org/) is an online repository of 3D structures of proteins, nucleic acids, and complex assemblies [34]. It currently holds 173 005 entries of protein structures. *Schistosoma*-related structures numbered 273 comprising 187 *Schistosoma mansoni*, 34 *Schistosoma japonicum*, 18 *Schistosoma haematobium,* and 34 others. These structures are essential in drug discovery.

SchistoDB (https://schistodb.net/schisto/app) is a database containing genomic data for blood flukes of the *Schistosoma* genus [35]. It contains whole-genome sequences, transcriptomic data, metabolic pathways, receptors, and other functionally enriched data. It also contains compounds assigned cHEBi and KEGG IDs which can be curated and deployed in the drug discovery process.

Table 1 Assays found in BindingDB on Schistosoma with information on ChEMBL ID, target or host receptor, and assaying techniques

ChEMBL ID	Target/Host Receptor	Technique/Assay
ChEMBL_589927	Thioredoxin glutathione reductase of *Schistosoma mansoni*	Functional assay
ChEMBL_1750450	Histone deacetylase 8 of *Schistosoma mansoni*	Fluorimetric assay
ChEMBL_1664931	Histone deacetylase 8 of *Schistosoma mansoni*	Microfluidic assay
ChEMBL_1742912	Histone deacetylase 8 of *Schistosoma mansoni*	ITC
ChEMBL_1455608	Thioredoxin glutathione reductase of *Schistosoma mansoni*	Functional assay
ChEMBL_1742907	Histone deacetylase 8 of *Schistosoma mansoni*	Fluorometric method
ChEMBL_1575026	Histone deacetylase 8 of *Schistosoma mansoni*	Fluorescence assay

Data characteristics

Data from these databases are available in various data formats based on various representative chemical notations such as *Simplified Molecular Input Line Entry System* (SMILES) and its variants as well as the Molfile family of formats (Chemical table files such as SDF, RGF, RDF and so on) [38]. These data representations are crucial to making chemical structures of molecules and atoms machine-readable for computational processes [31]. In terms of the representation of properties of said molecules, molecular or chemical descriptors are used. Molecular descriptors encode various physicochemical and topological properties that can then be utilized using various computational approaches (especially machine learning) to categorize or predict whether such properties make the small molecules potential candidates for activity against a given drug target [38]. For most machine learning purposes (some models can learn data representations via feature extraction without the need for descriptors), computation of these descriptors from the aforementioned structural representations is crucial and can directly influence results [39]. For a given molecule, hundreds to thousands of descriptors including fingerprints can be computed. For instance Mold2 can give about 777 descriptors for a single compound or small molecule [40]. PowerMV has been used to compute 179 descriptors for a biodataset of over 300 000 compounds screened against the *Schistosoma mansoni* thioredoxin glutathione reductase *(Sm*TGR*)*. Among these descriptors were properties such as polar surface area, hydrophobic bonds, molecular weight, the presence of charged groups, hydrogen-bond donors [23]. Detailed coverage of molecular descriptors including types and features is discussed elsewhere [41] [38]. However, within our scope, we point out their importance for small molecule prediction. Some popular software available for computing them include Mold2 (computes 777 descriptors) [40], Dragon 7 (5270 descriptors), Mordred (1825 descriptors), and PaDEL (1875 descriptors) [42].

In general, a biochemical dataset is obtained from a database, molecular descriptors are generated, and then used to train a suitable ML model via supervised learning [43]. Supervised learning enables the prediction of the label of new observations using an algorithm trained on a larger database of labeled examples [44]. The trained model can then be deployed in the drug discovery pipeline to predict hits which reduces compounds that need to be screened. The compounds and molecules in the dataset for training have been tested in wet-lab experiments for their inhibitory activity against a given disease target. Each is thus labeled active, inactive, or inconclusive based on their inhibitory activity using metrics such as their IC_{50} [45]. ML algorithms applied this way are also known as classifiers, and the most representative ones applied in drug discovery include Random Forest (RF), *Naive-Bayesian* (NB), support vector machine (SVM), and neural networks. However, there is no ideal classifier for all possible training sets: performance depends on the data [46]. So far, RF, NB and J48 are the ML classifiers used to predict antischistosomal compounds via binary classification [23]. Binary classification is used to predict antischistosomal compounds or molecules by classifying them under one of two categories: “active” or “inactive” against a given drug target [47]. This conversion of data to binary form is not without its challenges. The discrete nature of the categorization and cut-offs leads to the loss of insight into unique innate features that make interpretation more meaningful and representative of reality. This has led to the emerging trend of the use of continuous-valued datasets [48][49][43]. That notwithstanding, the binary strategy is the most developed aspect of learning from imbalanced biochemical data and is thus quite reliable [50].

Another noteworthy feature of datasets obtained from the aforementioned sources is that apart from missing data points, the data may suffer from a highly skewed distribution with one class having more representation and is thus the majority class. This is problematic for learning algorithms that favor the majority, especially when in most cases, the minority class is more important yet under-represented giving rise to data imbalance [50]. Data imbalance has a few variants discussed [50], but they all have the feature of heavy skewing of data to one class. For instance, the largest *Schistosoma*-related datasets (to authors’ best knowledge) were bioassays AID485364 and AID448 on PubChem (https://pubchem.ncbi.nlm.nih.gov/). The former had about 360 000 substances screened against the *Sm*TGR target with 10 784 actives compared to 334 477 inactives yielding a ratio of 1:30. AID448 had a low ratio with a staggering 1:1460 given that there were only 44 actives in the dataset of 64 651 substances. Such is the nature of most biochemical datasets available for drug discovery. It is thus difficult and sometimes near-impossible to use such data for machine learning without some interventions.

These interventions which tackle data imbalance can be categorized under three groups: data-level, algorithm-level and hybrid methods [51]. The first group is made up of techniques that modify the data to balance the distribution by adding synthetic samples to the minority class (oversampling) or reducing sample size of the majority class (undersampling) [50]. Examples are the Synthetic Minority Oversampling Technique (SMOTE) [15] and its proposed variants such as kSMOTE [52], and LN-SMOTE [53]. Other techniques that affect data only include the use of log-transforms and data imputation. The misuse of log-transforms, however, has led to the proposal of newer data analytic methods such as generalized estimation equations (GEE) [54]. Data imputation tackles the issue of missing data, its patterns and various methods to address it such as

spatial interpolation techniques [55]. In fact, DL has been used to perform missing data imputation in attention-deficit/hyperactivity disorder (ADHD) rating scales [56]. Algorithm-level methods modify existing algorithms such that they can adapt to skewed data and learning without being biased to the majority class. The main technique here is cost-sensitive learning where the algorithm is made to give more importance to the minority class via weight assignments or apply costs on misclassification [50, 57]. This method was used in training ML models for predicting *Sm*TGR inhibitors resulting in cost-sensitive classifiers [23]. The last group is the proposed hybridization of both data-level and algorithm-methods of which SPIDER is a typical example [51]. These interventions, however, give a slight drawback since boosting one class comes at a cost to their other hence, the introduction of data balancing techniques requires the use of additional scoring metrics to reflect the true performance of algorithms fitted to imbalanced data. A scoring metric such as overall accuracy is inadequate to depict the true performance for highly imbalanced data hence the use of better accuracy metrics such as balanced classification rate (BCR), Matthew's correlation coefficient (MCC), the geometric mean of sensitivity and specificity (G-means), the harmonic mean of sensitivity and specificity (F1-means) [14, 52]. The area under the curve (AUC) score for the receiver operating characteristic curve (ROC) also provides good insight into the predictive power of an ML model trained on imbalanced data, although it may have limitations [51].

Artificial intelligence (AI)

AI is a general term that implies the use of a computer to model intelligent behavior with the minimal human intervention [58]. AI and big data modeling have been applied in drug discovery [26, 59–61]. However, for this review, the sub-field of ML is the primary focus. In general, applications of AI in drug discovery and research include finding molecular targets, searching for hits or leads, synthesis of drug-like compounds, drug repurposing, and selection of a population for clinical trials [44–46]. AI can be applied at every stage of chemical design for drug discovery from the design and optimization of lead compounds to predicting absorption, distribution, metabolism and excretion, and toxicity (ADMET) properties [62].

With about only 10^8 synthesized compounds out of the estimated drug-like chemical space of approximately 10^{60}, there is the need to employ AI techniques to build autonomous experimental systems for drug discovery like Adam and Eve, which are robot scientists [63–66]. Eve performs compound screening, hit confirmation, and lead generation using QSAR models at a faster and more economical rate [65]. Eve was used to repurpose drugs for the treatment of NTDs including schistosomiasis [65]. These autonomous drug discovery systems can be leveraged to identify effective antischistosomal compounds by optimizing several drug properties simultaneously, taking into consideration the synthetic accessibility of the molecules.

Machine learning (ML)

ML is a sub-field of AI and consists of statistical and computer science principles to develop algorithms capable of improving performance through the interpretation of data rather than through explicit instructions [67]. Over the years, ML has been successfully used to extract knowledge from big data in bioinformatics [68–70]. ML algorithms use these datasets to uncover patterns, build models, and make predictions with the best model [24].

ML is sub-categorized into supervised, unsupervised, and reinforcement learning. Supervised learning involves the development of predictive models based on labeled data from both input and output sources. Such models aid in the diagnosis, prediction of inhibitors, and drug efficacy. Unsupervised learning algorithms group and interpret only input data to aid in disease target discovery among other applications. Reinforcement learning comprises decision-driven algorithms that execute tasks in a given environment and maximize their performance [47]. Comprehensive descriptions of ML categorizations and some drug-related applications are discussed elsewhere [45–48].

ML algorithms used in the prediction of antischistosomal molecules

Random forest (RF)

RF algorithm consists of several correlated decision trees as an ensemble. Each tree determines a prediction and the one that scores the most votes among the trees is given as the final prediction [47]. In other words, it is an ensemble of base classifiers called classification or decision trees. Technically, this makes RF a combined classifier that has the advantage of handling large datasets with superior accuracy [71]. Another popular classifier that also has a similar tree-based structure is the J48 algorithm from Weka [72].

Random forest was trained as a binary classifier on an imbalanced dataset of about 360 000 compounds with 10 784 biologically active ones against *Sm*TGR [23]. Since the data was imbalanced, cost-based learning was implemented with two methods: direct and meta learning. Suitable accuracy metrics such as MCC, AUC ROC score and BCR were used to evaluate the model performance. Random forest beat other models in the same study as the most sensitive classifier with AUC = 0.87, BCR = 80.1% and the highest MCC of 0.25. J48, followed closely with MCC = 0.23 and

BCR = 77.3% [23]. These models' predictions were validated and confirmed by methods such as molecular docking, thus supporting this binary approach [73].

Bayesian

Naive-Bayesian (NB) is the simplest form of probabilistic Bayesian classifier which assigns the most probable a posteriori class to a given instance based on Bayes' theorem with the assumption of independence between features [74]. NB had a sensitivity of 50.3% in predicting inhibitors of *Sm*TGR [23]. Despite this, NB has an advantageous ability to process vast data while handling random noise well [47].

Eight Bayesian models were trained on 3 898 data points from phenotypic screens against the schistosomula and adult stages of *S. mansoni* [75]. Two rule books were used to transform the data and then the models were trained according to the parasite's developmental stage and experimental time point ($\leq$24, 48, 72, and > 72 h). The models were used for activity predictions across commercial compound libraries resulting in 40 actives and 16 inactives. In vitro studies showed the prediction accuracies for active and inactives were 61% and 56% for somules and adults, respectively. The hit rates were thus 48% and 34%, respectively, far exceeding about 2% hit rate for traditional high-throughput screening [75] which also speaks well for binary classification.

Support vector machine (SVM)

SVM is a machine learning algorithm that separates classes of compounds based on a feature selector by deriving a hyperplane as a decision boundary. This hyperplane is derived by maximizing the margin between classes in N-dimensional space where N represents the number of features [76]. In training SVM, descriptors can be used as feature selectors. The hyperplane then divides compounds into positive in one direction and negative in another. The further away the compound is from the hyperplane, the less likely it will be selected. Thus, this algorithm can distinguish between active and inactive compounds.

SVM was one of many applied as binary QSAR models for virtual screening to obtain antischistosomal hits.The *Sm*TGR data the SVM was trained on were of two categories of balanced and unbalanced data in various ratios of active to inactive (1:2 and 1:3, respectively). The training combined AtomPair, molecular access system (MACCS), and Morgan fingerprints, chemistry development kit (CDK), and Dragon descriptors along with eight machine learning methods including RF, SVM and gradient boosting machine (GBM), resulting in 120 different QSAR model variants for both balanced and unbalanced data [5]. It was concluded that SVM, RF and GBM were the three best performing ML algorithms based on correct classification rate (CCR), Cohen's κ coefficient, sensitivity, and specificity. For instance, AtomPair-SVM, CDK-SVM, and Dragon-SVM had CCR scores of 0.81, 0.84, and 0.85, respectively. The hybrid models were combined to prioritize 29 compounds for further testing in two HCS platforms based on image analysis of assay plates resulting in 2 new antischistosomal hits [5].

Proposed deep learning paradigm for predicting antischistosomal small molecules

ML algorithms rely on data representations known as features. The needed expertise and difficultly of generating these features became a drawback to traditional ML algorithms until a new branch of ML emerged known as deep learning [77]. Deep learning uses artificial neural networks that adapt and learn from vast experimental data and thus overcomes some limitations of traditional ML [70]. Although, the deployment of ML techniques especially DL in drug development is not without its limitations, this approach has been used with varying successes. It is thus no surprise that DL methods are the new trend in modern drug discovery under the big data era [26]. Techniques such as high-throughput computational analysis of databases for lead and target discovery have paved the way for the incorporation of ML and DL techniques or algorithms into drug discovery [47].

Most deep learning algorithms have the structure of an artificial neural network (ANN) which comprises interconnected building blocks or units called artificial neurons that mimic the neurons in the central nervous system [77]. Therefore, the various types of deep learning can be categorized according to the arrangement and structure of these units (architecture). The main ones include *deep neural networks* (DNNs), *convolutional neural networks* (CNNs), *recurrent neural networks* (RNNs), and *emergent architectures* [24]. There are several other categorizations and types which have been discussed previously [78].

DNN algorithms are *multi-layer perceptrons* (MLPs), *stacked auto-encoders* (SAEs), and *deep belief networks* (DBNs) [24]. CNNs are similar to DNNs except that their units are arranged in a triple layer: convolution, non-linear, and pooling. GoogLeNet, AlexNet, and VGG architectures are examples of CNNs [24]. RNNs comprise cyclic connections of building blocks which are either *long short-term memory units* (LSTMs) or *gated recurrent units* (GRUs). Emergent architectures comprise hybrid algorithms that combine features of other architectures and algorithms. Examples are *deep Spatio-temporal neural networks* (DST-NNs), *multi-dimensional recurrent neural networks* (MD-RNNs), and *Bayesian regularized artificial neural networks* (BRANN) [24, 77, 79]. Although they are all applied across supervised learning, unsupervised learning and

representative learning, our focus will be on deep learning algorithms on supervised learning under which binary classification falls (Fig. 1).

Deep neural network (DNN)

This algorithm is based on the structure of ANNs which comprises an input layer, several hidden layers, and an output layer made up of interconnected units called nodes or neurons [52, 54]. The simplest structure of the DNN is a perceptron that separates two classes of data using a discrimination function; linear or non-linear. Examples are *sigmoid, rectified linear unit (ReLU), LeakyReLU,* and *softmax* [80, 81]. DNNs are further split into *multi-layer perceptrons* (MLPs), *stacked autoencoders* (SAEs), and *deep belief networks* (DBNs) based on the type of layers, units, and type of learning. The number of nodes and the number of hidden layers determine the network's width and depth respectively. This gives another categorization as those with several hidden layers containing a large number of nodes and those with a single hidden layer with a small number of nodes (shallow neural networks). Apart from this structural difference, weight initialization in the hidden layer nodes, transfer function applied and regularization could all differ. However, in terms of performance, the deeper networks have no significant advantage over their shallow counterparts [78, 82, 83]. DNNs in general can handle large raw data and also discover highly abstract patterns [84]. MLPs are used for supervised learning using labeled data, while SAEs and DBNs carry out unsupervised pre-training before supervised learning known as fine-tuning. SAEs and DBNs are made up of autoencoders and restricted Boltzmann machines respectively [24]. Each of these three algorithm subtypes can be modified to carry out binary classification on a given data of antischistosomal compounds. Deploying DNN for predicting antischistosomal molecules will be highly advantageous given that DNN outperforms other algorithms in various drug discovery studies [10, 24, 85].

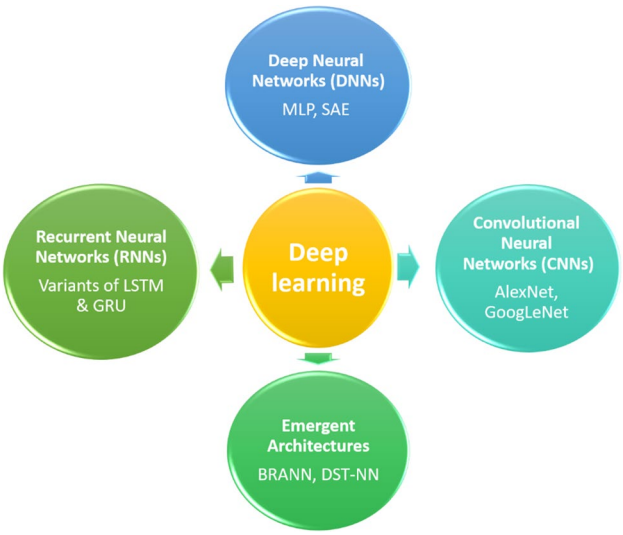


Fig. 1 The various types of deep learning algorithms to consider for drug discovery. Neural networks are categorized by their architectures into deep neural networks (DNNs), recurrent neural networks (RNN), convolutional neural networks (CNN), and hybrid emergent architectures

Limitations and recommendations

No single ML algorithm is suitable for all applications since each one has both advantages and disadvantages, many of which are data-dependent. The tree-based algorithms are highly sensitive, NB handles noise well, and SVM has a good binary classification capability [76]. DNN's ability to handle raw data and its self-learning capability also give it an edge in many applications such as feature extraction [86]. Hence, combining the capabilities of these ML algorithms in a multifaceted approach can generate the desired models to effectively reduce costs in drug discovery similar to the consensus models used by [5].

Algorithm limitations include underfitting and overfitting which both give inaccurate results. Overfitting is when the trained model learns so much that it captures irrelevant features of the data during the training which in turn impacts negatively on its performance [87]. Underfitting, on the other hand, is the trained model's inability to detect trends in the training data thereby making it unable to generalize well on new data [47]. Training algorithms are computationally tasking and intensive, and despite the advancements such as graphics processing units (GPU) computation, such technologies are not yet widespread [26].

Datasets used to train ML models are not devoid of issues as discussed earlier. This is significant since the accuracy of training data directly affects the model accuracy [88]. Another issue is limited data [28]: despite the astronomical growth of biodatasets for drug discovery, those of some neglected tropical diseases including schistosomiasis lags. This is a challenge for obtaining accurate predictive algorithms that need to learn from sufficient data [13]. This is further compounded by privacy restrictions and ethical requirements associated with data and by extension, big data [89]. Without sufficient datasets, it is difficult to externally evaluate the practical accuracy of the trained models [90]. Even when sufficient data is obtained, it is imbalanced such that the class distributions are very lopsided. A typical example is a study that had over 300 000 inactive compounds and just over 10 000 active ones [23]. These challenges limit the smooth training and deployment of accurate ML models.

A strategy composed of three techniques to tackle limited and imbalanced data for training ML models includes data

pre-processing, cost-sensitive learning, and algorithm modifications. Data pre-processing involves extracting redundant features and utilizing sampling methods discussed earlier (oversampling and undersampling) to balance the class distribution of data. A typical example of such a sampling method is the synthetic minority oversampling technique (SMOTE). Cost-sensitive learning is the application of costs for misclassification of data from certain classes. There are several explicit or implicit techniques to implement cost-sensitive learning, one of which is changing the objective loss function [24]. Algorithm modification comprises approaches that incorporate different learning algorithms to suit imbalanced and scarce data applications. Examples are unsupervised pre-training and transfer learning [91].

To overcome data-related limitations, worldwide public–private partnerships (PPPs) were proposed to advance research infrastructure about technologies, facilities, and human resources for the schistosomiasis drug development pipeline [6]. Exploiting such a networking initiative can alleviate the scarcity of specific data and improve global accessibility [6]. Also, the disjoint among different scientific groups should be bridged by collaborative efforts through consortia and social media [13]. This can facilitate sharing of data, computational models, and cross-field collaborations. Also, schistosomiasis-specific databases containing datasets curated from experimental work and those scattered in a myriad of databases can be developed as a one-stop shop for harvesting antischistosomal molecules for use in ML algorithm developments. Such a database can contain datasets that have been properly formatted and assigned ontological features to enable a multifaceted approach to the design of ML models.

Conclusion

The big data paradigm provides an opportunity to fully leverage ML algorithms to reduce costs for drug discovery in schistosomiasis with the availability of ever-expanding datasets on open databases. Algorithms such as SVM, RF, J48, and NB have already been deployed as binary classifiers to predict antischistosomal molecules or compounds. Deep learning algorithms such as DNN-based MLP, DBN, and SAE are also potentially useful in this application. Combining various algorithms is another potential approach that can be adopted. By increasing collaborative efforts among stakeholders and adapting efficient ML training approaches, the limitations to full exploitation of these ML approaches in the drug discovery pipeline can be reduced. This, in turn, can speed up the search for efficacious therapeutics to tackle this neglected tropical disease. Besides, the availability of bioactive datasets of antischistosomal molecules obtained from HTS experiments is a useful resource for developing efficient and robust ML models.

Authors' contributions SKK conceptualized the review. SKK and KAM co-wrote the first draft with contributions from WAM, EB and MDW. All authors read and accepted the final draft for submission.

Funding The work was not funded.

Declarations

Conflict of interest The authors declare no conflicts of interest.

References

1. LoVerde PT (2019) Schistosomiasis. Advances in Experimental Medicine and Biology. Springer, New York LLC, pp 45–70
2. Adenowo AF, Oyinloye BE, Ogunyinka BI, Kappo AP (2015) Impact of human schistosomiasis in sub-Saharan Africa. Braz J Infect Dis 19:196–205. https://doi.org/10.1016/j.bjid.2014.11.004
3. Vos T, Abajobir AA, Abbafati C et al (2017) Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016. Lancet 390:1211–1259. https://doi.org/10.1016/S0140-6736(17)32154-2
4. Freer JB, Bourke CD, Durhuus GH et al (2018) Schistosomiasis in the first 1000 days. Lancet Infect Dis 18:e193–e203
5. Neves BJ, Dantas RF, Senger MR et al (2016) Discovery of new anti-schistosomal hits by integration of QSAR-based virtual screening and high content screening. J Med Chem 59:7075–7088. https://doi.org/10.1021/acs.jmedchem.5b02038
6. Moreira-Filho JT, Dantas RF, Senger MR, et al (2019) Shortcuts to schistosomiasis drug discovery: The state-of-the-art. In: Annual Reports in Medicinal Chemistry. Academic Press Inc., pp 139–180
7. da Siqueira L, P, Fontes DAF, Aguilera CSB, et al (2017) Schistosomiasis: drugs used and treatment strategies. Acta Trop 176:179–187
8. Bergquist R, Elmorshedy H (2018) Artemether and praziquantel: Origin, mode of action, impact, and suggested application for effective control of human schistosomiasis. Trop. Med. Infect. Dis. 3
9. Tavares NC, de Aguiar PHN, Gava SG, et al (2016) Schistosomiasis: Setting Routes for Drug Discovery. In: Special Topics in Drug Discovery. InTech
10. Xu JF, Xu J, Li SZ et al (2013) Transmission risks of schistosomiasis japonica: extraction from back-propagation artificial neural network and logistic regression model. PLoS Negl Trop Dis. https://doi.org/10.1371/journal.pntd.0002123
11. Caffrey CR, Secor WE (2011) Schistosomiasis: from drug deployment to drug development. Curr Opin Infect Dis 24:410–417
12. Gouveia M, Brindley P, Gärtner F et al (2018) Drug repurposing for schistosomiasis: combinations of drugs or biomolecules. Pharmaceuticals 11:15. https://doi.org/10.3390/ph11010015
13. Ponder EL, Freundlich JS, Sarker M, Ekins S (2014) Computational models for neglected diseases: gaps and opportunities. Pharm Res 31:271–277. https://doi.org/10.1007/s11095-013-1170-9

14. Winkler DA (2021) Use of artificial intelligence and machine learning for discovery of drugs for neglected tropical diseases. Front Chem 9:1–15. https://doi.org/10.3389/fchem.2021.614073
15. Fusco T, Bi Y, Wang H, Browne F (2020) Data mining and machine learning approaches for prediction modelling of schistosomiasis disease vectors: epidemic disease prediction modelling. Int J Mach Learn Cybern 11:1159–1178. https://doi.org/10.1007/s13042-019-01029-x
16. Shen Y, Sung MH, King CH et al (2020) Modeling approaches to predicting persistent hotspots in score studies for gaining control of schistosomiasis Mansoni in Kenya and Tanzania. J Infect Dis 221:796–803. https://doi.org/10.1093/infdis/jiz529
17. Li G, Zhou X, Liu J et al (2018) Comparison of three data mining models for prediction of advanced schistosomiasis prognosis in the Hubei province. PLoS Negl Trop Dis. https://doi.org/10.1371/journal.pntd.0006262
18. Holmström O, Linder N, Ngasala B et al (2017) Point-of-care mobile digital microscopy and deep learning for the detection of soil-transmitted helminths and Schistosoma haematobium. Glob Health Action. https://doi.org/10.1080/16549716.2017.1337325
19. Angela MU, Oluwatosin AM (2016) Predicting the Risk of Infection with SCHISTOSOMA HAEMATOBIUM using Machine Learning
20. Garcia FP, Guedes GP, Belloze KT (2020) Identifying *Schistosoma mansoni* essential protein candidates based on machine learning. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer, pp 123–128
21. Campos TDL, Young ND, Korhonen PK, et al (2014) Identification of G protein-coupled receptors in *Schistosoma haematobium* and *S. mansoni* by comparative genomics. Parasit Vectors 7: 242. https://doi.org/10.1186/1756-3305-7-242
22. Rojo-Arreola L, Long T, Asarnow D et al (2014) Chemical and genetic validation of the statin drug target to treat the helminth disease. Schistosomiasis. https://doi.org/10.1371/journal.pone.0087594
23. Gaba S, Jamal S, Drug Discovery Consortium OS, Scaria V (2014) Cheminformatics models for inhibitors of *Schistosoma mansoni* Thioredoxin glutathione reductase. Sci World J 2014:1–9. https://doi.org/10.1155/2014/957107
24. Min S, Lee B, Yoon S (2017) Deep learning in bioinformatics. Brief Bioinform 18:851–869. https://doi.org/10.1093/bib/bbw068
25. Jin X, Wah BW, Cheng X, Wang Y (2015) Significance and challenges of big data research. Big Data Res 2:59–64. https://doi.org/10.1016/j.bdr.2015.01.006
26. Zhu H (2020) Big data and artificial intelligence modeling for drug discovery. Annu Rev Pharmacol Toxicol 60:573–589
27. Baro E, Degoul S, Beuscart R, Chazard E (2015) Toward a literature-driven definition of big data in healthcare. Biomed Res Int. https://doi.org/10.1155/2015/639021
28. Zhao L, Ciallella HL, Aleksunes LM, Zhu H (2020) Advancing computer-aided drug discovery (CADD) by big data and data-driven machine learning modeling. Drug Discov Today 25:1624–1638. https://doi.org/10.1016/j.drudis.2020.07.005
29. Lo YC, Rensi SE, Torng W, Altman RB (2018) Machine learning in chemoinformatics and drug discovery. Drug Discov Today 23:1538–1546
30. Kim S (2016) Getting the most out of PubChem for virtual screening. Expert Opin Drug Discov 11:843–855
31. Gaulton A, Bellis LJ, Bento AP et al (2012) ChEMBL: a large-scale bioactivity database for drug discovery. Nucleic Acids Res 40:D1100–D1107. https://doi.org/10.1093/nar/gkr777
32. Wishart DS, Feunang YD, Guo AC et al (2018) DrugBank 5.0: a major update to the drugbank database for 2018. Nucleic Acids Res 46:D1074–D1082. https://doi.org/10.1093/nar/gkx1037
33. Gilson MK, Liu T, Baitaluk M et al (2016) BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology. Nucleic Acids Res 44:D1045–D1053. https://doi.org/10.1093/nar/gkv1072
34. Burley SK, Berman HM, Kleywegt GJ, et al (2017) Protein Data Bank (PDB): The single global macromolecular structure archive. In: Methods in Molecular Biology. Humana Press Inc., pp 627–641
35. Zerlotini A, Aguiar ERGR, Yu F et al (2013) SchistoDB: An updated genome resource for the three key schistosomes of humans. Nucleic Acids Res. https://doi.org/10.1093/nar/gks1087
36. Kim S, Chen J, Cheng T et al (2021) PubChem in 2021: new data content and improved web interfaces. Nucleic Acids Res 49:D1388–D1395. https://doi.org/10.1093/nar/gkaa971
37. Mendez D, Gaulton A, Bento AP et al (2019) ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res 47:D930–D940. https://doi.org/10.1093/nar/gky1075
38. David L, Thakkar A, Mercado R, Engkvist O (2020) Molecular representations in AI-driven drug discovery: a review and practical guide. J Cheminform 12:1–22. https://doi.org/10.1186/s13321-020-00460-5
39. Kim H, Kim E, Lee I et al (2020) Artificial intelligence in drug discovery: a comprehensive review of data-driven and machine learning approaches. Biotechnol Bioprocess Eng 25:895–930. https://doi.org/10.1007/s12257-020-0049-y
40. Hong H, Xie Q, Ge W et al (2008) Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics. J Chem Inf Model 48:1337–1344. https://doi.org/10.1021/ci800038f
41. Ponzoni I, Sebastián-Pérez V, Requena-Triguero C et al (2017) Hybridizing feature selection and feature learning approaches in QSAR modeling for drug discovery /631/114/2248 /631/154/309 /639/638/563/606 /119/118 article. Sci Rep 7:1–19. https://doi.org/10.1038/s41598-017-02114-3
42. Moriwaki H, Tian Y-S, Kawashita N, Takagi T (2018) Mordred: a molecular descriptor calculator. J Cheminform 10:4. https://doi.org/10.1186/s13321-018-0258-y
43. Krstajic D, Buturovic L, Thomas S, Leahy DE (2017) Binary classification models with "Uncertain" predictions
44. Uddin S, Khan A, Hossain ME, Moni MA (2019) Comparing different supervised machine learning algorithms for disease prediction. BMC Med Inform Decis Mak 19:281. https://doi.org/10.1186/s12911-019-1004-8
45. Armutlu P, Ozdemir ME, Uney-Yuksektepe F et al (2008) Classification of drug molecules considering their IC50 values using mixed-integer linear programming based hyper-boxes method. BMC Bioinformatics 9:411. https://doi.org/10.1186/1471-2105-9-411
46. Uçar MK, Nour M, Sindi H, Polat K (2020) The effect of training and testing process on machine learning in biomedical datasets. Math Probl Eng. https://doi.org/10.1155/2020/2836236
47. Patel L, Shukla T, Huang X et al (2020) Machine learning methods in drug discovery. Molecules 25:5277. https://doi.org/10.3390/molecules25225277
48. Schmitz S, Adams R, Walsh C (2012) The use of continuous data versus binary data in MTC models: a case study in rheumatoid arthritis. BMC Med Res Methodol. https://doi.org/10.1186/1471-2288-12-167
49. Bagherian M, Sabeti E, Wang K et al (2021) Machine learning approaches and databases for prediction of drug-target interaction: a survey paper. Brief Bioinform 22:247–269. https://doi.org/10.1093/bib/bbz157
50. Krawczyk B (2016) Learning from imbalanced data: open challenges and future directions. Prog Artif Intell 5:221–232. https://doi.org/10.1007/s13748-016-0094-0

51. Stefanowski J Dealing with Data Difficulty Factors while Learning from Imbalanced Data
52. Raj KP, Raju KVS (2017) Using Machine Learning Algorithms To. 1:2007
53. Lago EM, Xavier RP, Teixeira TR et al (2018) Antischistosomal agents: state of art and perspectives. Future Med Chem 10:89–120. https://doi.org/10.4155/fmc-2017-0112
54. Feng C, Wang H, Lu N et al (2014) Log-transformation and its implications for data analysis. Shanghai Arch Psychiatry 26:105–109. https://doi.org/10.3969/j.issn.1002-0829.2014.02
55. Richman MB, Trafalis TB, Adrianto I (2009) Missing data imputation through machine learning algorithms. In: Artificial Intelligence Methods in the Environmental Sciences. Eds: Sue Ellen Haupt, Antonello Pasini, Caren Marzban. Springer Netherlands, pp 153–169
56. Cheng CY, Tseng WL, Chang CF et al (2020) A deep learning approach for missing data imputation of rating scales assessing attention-deficit hyperactivity disorder. Front Psychiatry 11:673. https://doi.org/10.3389/fpsyt.2020.00673
57. Zhou ZH, Liu XY (2010) On multi-class cost-sensitive learning. Comput Intell 26:232–257. https://doi.org/10.1111/j.1467-8640.2010.00358.x
58. Hamet P, Tremblay J (2017) Artificial intelligence in medicine. Metabolism 69:S36–S40. https://doi.org/10.1016/j.metabol.2017.01.011
59. de Jong J, Cutcutache I, Page M et al (2021) Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. Brain. https://doi.org/10.1093/brain/awab108
60. Keshavarzi Arshadi A, Webb J, Salem M et al (2020) Artificial intelligence for COVID-19 drug discovery and vaccine development. Front Artif Intell. https://doi.org/10.3389/frai.2020.00065
61. Mak KK, Pichika MR (2019) Artificial intelligence in drug development: present status and future prospects. Drug Discov Today 24:773–780. https://doi.org/10.1016/j.drudis.2018.11.014
62. Bruno S, Pharmaceutical T, Healthcare GNS, et al (2017) AI-powered drug discovery captures pharma interest. 35: https://doi.org/10.1038/nature22322
63. Kim S, Thiessen PA, Bolton EE et al (2016) PubChem substance and compound databases. Nucleic Acids Res 44:D1202–D1213. https://doi.org/10.1093/nar/gkv951
64. Polishchuk PG, Madzhidov TI, Varnek A (2013) Estimation of the size of drug-like chemical space based on GDB-17 data. J Comput Aided Mol Des. https://doi.org/10.1007/s10822-013-9672-4
65. Williams K, Bilsland E, Sparkes A et al (2015) Cheaper faster drug development validated by the repositioning of drugs against neglected tropical diseases. J R Soc Interface. https://doi.org/10.1098/rsif.2014.1289
66. Sparkes A, Aubrey W, Byrne E et al (2010) Towards Robot Scientists for autonomous scientific discovery. Autom Exp 2:1. https://doi.org/10.1186/1759-4499-2-1
67. Abbasi B, Goldenholz DM (2019) Machine learning applications in epilepsy. Epilepsia 60:2037–2047
68. Vamathevan J, Clark D, Czodrowski P et al (2019) Applications of machine learning in drug discovery and development. Nat Rev Drug Discov 18:463–477. https://doi.org/10.1038/s41573-019-0024-5
69. Larrañaga P, Calvo B, Santana R et al (2006) Machine learning in bioinformatics. Brief Bioinform 7:86–112
70. Nguyen G, Dlugolinsky S, Bobák M et al (2019) Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey. Artif Intell Rev 52:77–124. https://doi.org/10.1007/s10462-018-09679-z
71. Fawagreh K, Gaber MM, Elyan E (2014) Random forests: from early developments to recent advancements. Syst Sci Control Eng 2:602–609. https://doi.org/10.1080/21642583.2014.956265
72. Mahesh JU, Naganjaneyulu KV, Likitha P, Aishwarya KNSS (2014) Analysis of J48 algorithm in classification-ebola virus. Int J Emerg Trends Sci Technol 1:1289–1292
73. Peña-Guerrero J, Nguewa PA, García-Sosa AT (2021) Machine learning, artificial intelligence, and data science breaking into drug design and neglected diseases. Wiley Interdiscip Rev Comput Mol Sci. https://doi.org/10.1002/wcms.1513
74. Zhang Z (2016) Naïve bayes classification in R. Ann Transl Med. 4: 1–5. https://doi.org/10.21037/atm.2016.03.38
75. Zorn KM, Sun S, McConnon CL et al (2021) A Machine learning strategy for drug discovery identifies anti-schistosomal small molecules. ACS Infect Dis 7:406–420. https://doi.org/10.1021/acsinfecdis.0c00754
76. Cervantes J, Garcia-Lamont F, Rodríguez-Mazahua L, Lopez A (2020) A comprehensive survey on support vector machine classification: applications, challenges and trends. Neurocomputing 408:189–215. https://doi.org/10.1016/j.neucom.2019.10.118
77. Jing Y, Bian Y, Hu Z et al (2018) Deep learning for drug design: an artificial intelligence paradigm for drug discovery in the big data era. AAPS J. https://doi.org/10.1208/s12248-018-0210-0
78. Shrestha A, Mahmood A (2019) Review of deep learning algorithms and architectures. IEEE Access 7:53040–53065
79. Yamashita R, Nishio M, Do RKG, Togashi K (2018) Convolutional neural networks: an overview and application in radiology. Insights Imaging 9:611–629
80. Koutsoukas A, Monaghan KJ, Li X, Huan J (2017) Deep-learning: investigating deep neural networks hyper-parameters and comparison of performance to shallow methods for modeling bioactivity data. J Cheminform 9:42. https://doi.org/10.1186/s13321-017-0226-y
81. Nwankpa C, Ijomah W, Gachagan A, Marshall S (2018) Activation Functions: Comparison of trends in Practice and Research for Deep Learning. arXiv
82. Winkler DA, Le TC (2017) Performance of deep and shallow neural networks, the universal approximation theorem, activity cliffs, and QSAR. Mol Inform 36:1–6. https://doi.org/10.1002/minf.201600118
83. Lusci A, Pollastri G, Baldi P (2013) Deep architectures and deep learning in chemoinformatics: the prediction of aqueous solubility for drug-like molecules. J Chem Inf Model 53:1563–1575. https://doi.org/10.1021/ci400187y
84. Mahmud M, Shamim Kaiser M, Hussain A et al (2018) of deep learning and reinforcement learning to biological data. IEEE Trans Neural Netw Learn Syst. https://doi.org/10.1109/TNNLS.2018.2790388
85. Lenselink EB, Ten Dijke N, Bongers B et al (2017) Beyond the hype: deep neural networks outperform established methods using a ChEMBL bioactivity benchmark set. J Cheminform. https://doi.org/10.1186/s13321-017-0232-0
86. Bengio Y, Courville A, Vincent P (2013) Representation Learning : A Review and New Perspectives 35:1798–1828
87. Ying X (2019) An Overview of Overfitting and its Solutions. In: Journal of Physics: Conference Series. Institute of Physics Publishing
88. Lei S, Zhang H, Wang K, Su Z (2018) How training data affect the accuracy and robustness of neural networks for image classification
89. Kuc-Czarnecka M, Olczyk M (2020) How ethics combine with big data: a bibliometric analysis. Humanit Soc Sci Commun 7:1–9. https://doi.org/10.1057/s41599-020-00638-0
90. Mafud AC, Ferreira LG, Mascarenhas YP et al (2016) Discovery of Novel Antischistosomal Agents by Molecular Modeling

Approaches. Trends Parasitol 32:874–886. https://doi.org/10.1016/j.pt.2016.08.002
91. Cai C, Wang S, Xu Y et al (2020) Transfer Learning for Drug Discovery. J Med Chem 63:8683–8694. https://doi.org/10.1021/acs.jmedchem.9b02147

Authors and Affiliations

Samuel K. Kwofie[1,2] · Kwasi Agyenkwa-Mawuli[1,2] · Emmanuel Broni[1,6] · Whelton A. Miller III[3,4,5] · Michael D. Wilson[3,6]

✉ Samuel K. Kwofie
skkwofie@ug.edu.gh

[1] Department of Biomedical Engineering, School of Engineering Sciences, College of Basic and Applied Sciences, University of Ghana, Legon, P.O. Box LG 77, Accra, Ghana

[2] West African Centre for Cell Biology of Infectious Pathogens, Department of Biochemistry, Cell and Molecular Biology, University of Ghana, Legon, P.O. Box LG 54, Accra, Ghana

[3] Department of Medicine, Loyola University Medical Center, Maywood, IL 60153, USA

[4] Department of Molecular Pharmacology and Neuroscience, Loyola University Medical Center, Maywood, IL 60153, USA

[5] Department of Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104, USA

[6] Department of Parasitology, Noguchi Memorial Institute for Medical Research, University of Ghana, Legon, P.O. Box LG 581, Accra, Ghana