We are IntechOpen, the world's leading publisher of Open Access books Built by scientists, for scientists

4,800
Open access books available

122,000
International authors and editors

135M
Downloads

Our authors are among the

154
Countries delivered to

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

CLARIVATE ANALYTICS BOOK CITATION INDEX INDEXED

WEB OF SCIENCE™

Selection of our books indexed in the Book Citation Index in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Chapter

Designing Data-Driven Learning Algorithms: A Necessity to Ensure Effective Post-Genomic Medicine and Biomedical Research

Gaston K. Mazandu, Irene Kyomugisha, Ephifania Geza, Milaine Seuneu, Bubacarr Bah and Emile R. Chimusa

Abstract

Advances in sequencing technology have significantly contributed to shaping the area of genetics and enabled the identification of genetic variants associated with complex traits through genome-wide association studies. This has provided insights into genetic medicine, in which case, genetic factors influence variability in disease and treatment outcomes. On the other side, the missing or hidden heritability has suggested that the host quality of life and other environmental factors may also influence differences in disease risk and drug/treatment responses in genomic medicine, and orient biomedical research, even though this may be highly constrained by genetic capabilities. It is expected that combining these different factors can yield a paradigm-shift of personalized medicine and lead to a more effective medical treatment. With existing "big data" initiatives and high-performance computing infrastructures, there is a need for data-driven learning algorithms and models that enable the selection and prioritization of relevant genetic variants (post-genomic medicine) and trigger effective translation into clinical practice. In this chapter, we survey and discuss existing machine learning algorithms and post-genomic analysis models supporting the process of identifying valuable markers.

Keywords: learning algorithms, machine learning, genome-wide association study, genomic medicine, biomedical research, post-genomic analysis

1. Introduction

Advancements in the human deoxyribonucleic acid (DNA) microarray and genome sequencing technology have resulted in an exponential growth of publicly available and accessible biological datasets [1, 2]. These "big data" are being explored to systematically uncover useful signals and gain more insights to advance current knowledge and answer specific biological and health questions. Considering current data delude and relatively increased computing power, it is becoming possible to accurately infer desirable features from such data. This highlights the need for efficient learning algorithms to process these data for knowledge discovery by identifying pertinent patterns related to the comparison and classification of different features in these datasets. These learning algorithms should enable

the extraction of appropriate features for application in a novel event or situation to support decision-making by mapping a given system to an input-output transformation task as shown in **Figure 1**. Emerging trends in (deep) machine learning algorithms have made possible the identification and discovery of new patterns and hidden processes in genomic sequences that are essential in the functioning of biological systems. The heterogeneity of diseases, such as cancer, requires primarily the analysis of genomic data in order to improve diagnosis and to design an optimal therapy for an efficient clinical management of the disease. There is an increasing need of machine learning techniques in genomic medicine.

Machine learning algorithms can be classified into three main categories, namely *supervised*, *unsupervised* and *reinforcement* learning, described below:

Supervised learning algorithms build a mapping function, f, from the input variable, X, to the output result, Y, expressed by: $Y = f(X)$. There exist two main groups of supervised learning algorithms, namely classification and regression. Classification model is used to predict the outcome of a given sample with categorical output, for instance, case or sick individuals, labeled 0, and control or healthy individuals, labeled 1. On the other hand, a regression model is used to predict the outcome of a given sample with a real-valued output. Examples of supervised learning algorithms include logistic and linear regression models, Naive Bayes, classification and regression trees (CART) [3], K-nearest neighbor (KNN) [4, 5], support vector machine (SVM) [6], random forest (RF) [7], and artificial neural networks (ANNs) [8].

Unsupervised learning algorithms retrieve the underlying structure of the dataset based on input X only, using unlabeled data, that is, input data with no corresponding output. In this type of learning algorithm, we have: *clustering*, *dimensionality reduction*, and *association* models. Clustering consists of grouping samples so that items within the same cluster are more similar to each other than to items from another cluster for a given well-defined metric. Dimensionality reduction uses feature extraction and selection methods to reduce the number of input variables, conveying the most important information and minimizing noise in the dataset. Feature selection extracts a subset of useful variables among the original variables and transforms data from a high- to a low-dimensional space. Finally, association model just computes the probability of the co-occurrence of elements in

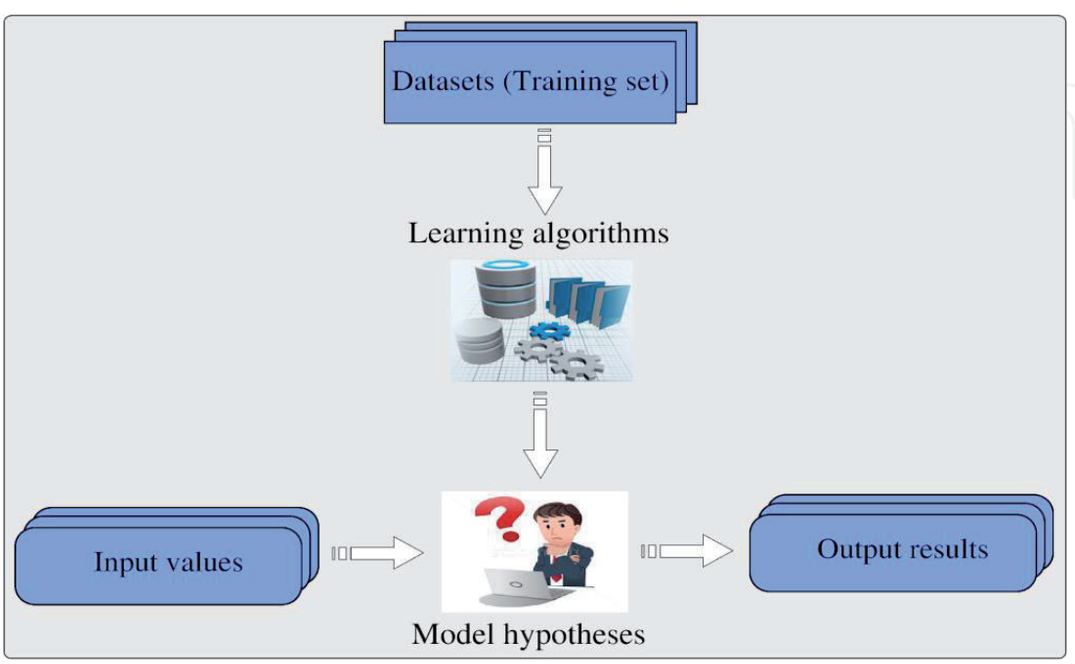

Figure 1.
Mapping a system to an input-output transformation task through learning algorithms namely supervised, unsupervised, and reinforcement learning.

a collection, thus inferring how likely two different elements are to co-occur in a collection. Unsupervised learning includes hierarchical clustering [9], K-means [10], and principal component analysis (PCA).

Reinforcement learning algorithms are a class of learning algorithms allowing an agent to decide the optimal next action based on its current state to control an environment or a system [11], by learning behaviors that will maximize the reward or outcome [12, 13]. These algorithms interact with a system, for example human system under a specific condition which may be disease or treatment, to learn the best setting and optimally perform sequential decisions along a timeline [11], generally under uncertainty, based solely on the present state of the system. It follows that this sequential and dynamic decision-making process is assumed to be a Markov decision process [14], in which the present state of the system fully describes the system and is sufficient to optimally predict the best next state. Reinforcement learning algorithms generally use a dynamic programming method following Bellman-based optimality principle [12], which requires optimal substructure for a given optimal option. In clinical research, these algorithms can be effective for longitudinal analyses, including retrospective and prospective studies, which consist of following a cohort across a specific-time interval [15].

Most of these learning algorithms have been extensively used to overcome several issues in genomic medicine, including identification of genetic markers underlying disease risk, novel mechanisms for disease prevention, control, diagnosis and therapy, building predictive disease models, predicting treatment outcomes, etc. Currently, there exist several platforms producing large-scale datasets, including genomics, transcriptomics, proteomics, metabolomics, and microbial and epidemiological data. This provides a unique opportunity of setting models and learning algorithms to enable the integration of these different heterogeneous datasets for elucidating determinant factors contributing to disease outcome and therapy in order to take full advantage of this data wealth in post-genomic medicine. In the following sections, we review some cases where machine and deep learning techniques have been used in health era and how post-genomic analysis constitutes a necessary route for optimally elucidating mechanisms of disease for an appropriate disease clinical management, and for predicting effective therapeutic strategies.

2. Use of machine learning in biology and health domains

As pointed out previously, machine learning algorithms have been successfully applied in many areas of biology and health-related research, including the identification of previously unknown processes in the genome, identification and understanding of several differentially expressed genes, binding specificities, and alternative splicing effects on cell processes, gene-gene and gene-environment interactions, disease-causing mutations, genetic determinants of diseases, pathway analysis, network and co-expression analysis, prediction of new drug-targets and response to treatment, etc. Here, we provide some illustrations of the use of supervised classification machine learning algorithms such as regression, SVM, ANN, and RF in some specific genomic applications, including predicting sequence specificities, analyzing gene expression profiles, identifying gene-gene and protein-protein interactions, and elucidating disease-associated variants.

2.1 Predicting sequence specificities of DNA- and RNA-binding proteins

Sequence specificities of DNA- and RNA-binding proteins are essential for developing models of regulatory processes in biological systems. Alipanahi et al. [16] present the possibility of predicting sequence specificities from experimental data through

deep learning. They developed a software tool (DeepBind) based on deep convolutional neural networks that has the ability to discover new patterns in a sequence without knowledge of the particular location of the pattern within the sequence. DeepBind is also said to have the ability to: learn from very large amounts of sequence data through parallel implementation on a graphics processing unit (GPU); use both microarray and sequencing data; automatically train predictive models without requiring hand-tuning; tolerate mislabeled data and some noise; and generalize well across technologies regardless of existing biases across technologies. Furthermore, DeepBind was also used for identifying RNA- and DNA-binding protein sequence specificities, and showed resilience to outliers and array biases. This suggests that the issue of predicting sequence specificities has been efficiently addressed using the deep learning approach.

2.2 Analyzing gene expression profiles

With the increased availability of genome-wide gene expression assays in public databases, there is increasing demand for more efficient computational models for data interpretation. The use of artificial neural networks in biomedical research is currently taking precedence over traditional analysis methods, as they have been proven to be better classifiers. Deep neural networks, using data from RNA-seq as inputs, are being used for prediction modeling. Classic models in applications like predicting patient outcomes using gene expression data are still not effective to the expected level, thus creating a need for more efficient robust algorithms. Recent studies that use deep learning models on gene expression data have indicated better performance. Urda et al. [17] illustrated the use of a multi-layer feed-forward artificial neural network, shown in **Figure 2**, in analyzing the RNA-seq gene expression data.

Dincer et al. [18] present a model that uses variational auto-encoders (VAEs) to extract latent variables from publicly available expression datasets and use them as features for predicting phenotypes. Their system, called DeepProfile, uses deep learning to learn a feature representation from large unlabeled expression data samples that are not incorporated in the prediction problem. This system was successfully used for the prediction of response to cancer drugs based on gene expression data. It also helped determine the effects of given drugs on specific patients

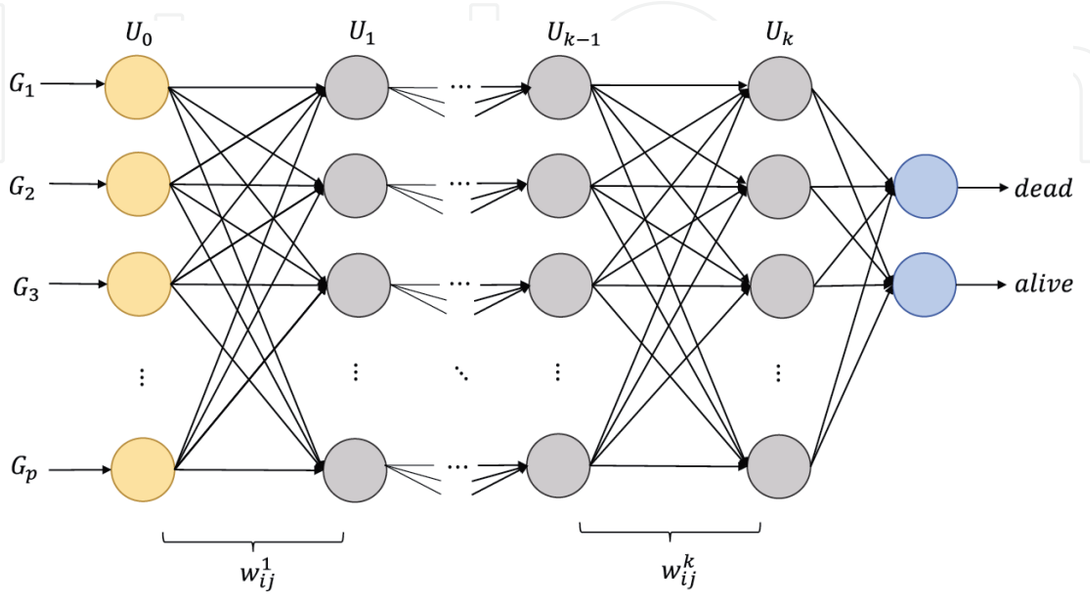

Figure 2.
Example neural network for binary classification. Input layer of P gene expression levels connected to k-hidden layers through synaptic weights w.

and thus provides a tool for precision medicine. The model was trained on gene expression data of acute myeloid leukemia, from GEO. Results indicated that low-dimensional representation (latent variables) generated using VAEs significantly outperformed the original input feature representation (gene expression levels) in the drug response prediction problem. Therefore, variational auto-encoders were shown to be effective in extracting a low-dimensional feature representation from unlabeled gene expression datasets and these learned features were found to capture important processes relevant to the prediction problem.

It is worth noting that detecting certain differentially expressed genes (DEGs) from RNA-seq results still faces challenges despite the quality control measures applied during sample preparation and data analysis. Data processing methods can lead to a certain number of false-positives and false-negatives that affect the accuracy and sensitivity of DEGs analysis. The combination of machine learning techniques with RNA-seq has been shown to significantly improve the sensitivity of DEGs [18] and thus help increase the identification of DEGs that are missed by traditional RNA-seq techniques. The study by Wang et al. [19] used a differential network analysis, based on machine learning, to predict stress-responsive genes by learning the patterns of 32 expression characteristics of known stress-related genes. For analysis using machine learning, the WEKA 3 data mining software was used for feature selection, classifier training, and evaluation. Three feature selection algorithms, correlation feature selection (CFS), information gain (InfoGain), and RELIEF [20], were used to identify features and five classifiers, logistic regression, random forest, LMT, classification via regression, and random subspace, that exhibited better performance than other machine learning algorithms, were deployed to predict up- and down-regulated genes. With this approach, the authors were able to identify the top 23 most informative features.

2.3 Inferring protein-protein interaction and biological networks for knowledge discovery

In the context of this chapter, we only focus on protein-protein interaction (PPI) network, which is defined as a set of nodes (or vertices), representing proteins connected by undirected edges (or links), which are the interactions or relationships between them (either *direct physical* or *functional* interactions). A physical interaction is an interaction that involves physical contact between proteins, and on the other hand, functional interaction, which is broad, does not necessarily involve direct physical contact, but rather refers to a mechanism through which a protein participates in cell functions [21]. Several learning algorithms have been used to infer human and human-pathogen PPIs [22], including ANN [23].

There exist several types of PPI networks based on the type of interactions and when integrated in a single network, the relationships between proteins in a unified network are referred to as functional interactions. Here, we only refer to functional interactions, which include physical and genetic interactions, and those inferred from knowledge about co-expression and shared evolutionary history or biological pathways. Other types of biological networks include signaling networks, gene regulatory or DNA-protein interaction networks [24, 25], disease-gene networks linking diseases to genes causing the disease, and drug interaction networks connecting drugs to their targets [26]. These biological networks have been used in several applications and analyzing individual, collective, and sub-network behaviors of these biological networks has enabled effective knowledge discovery at different levels of biology.

2.4 Predicting gene-gene and gene-environment interactions

Generally, disease outcome involves multiple genes contributing in every stage of disease progression [27]. This suggests the influence of gene-gene and gene-environment interactions in the outcome of a disease. Genes interact in large networks and some genes in the network are more important or central than others. Understanding these interactions is necessary for setting optimal prevention and control mechanisms to contain the disease. There have been challenges in identifying the distinctive nature of gene-gene and gene-environment interactions and their impact on disease risk, using traditional statistical methods. This has been due to the high dimensionality of the data, presence of epistasis and multiple polymorphisms leading to complex datasets for analysis. Machine learning methods such as SVM, ANN, and RF are used in addressing these challenges.

Neural networks use pattern recognition to address challenges in genomics. In the context of predicting gene-gene interaction, the neural network architecture depends on the type of interactions [28], shown in **Figure 3**. Genetic programming has been utilized to optimize the architecture of neural networks and back propagation for modeling gene-gene interactions as illustrated by Ritchie et al. [29]. Genetic programing neural nets (GPNN) were found to have more prediction power for models with heritability greater than 0.026 as compared to back propagation neural nets (BPNNs) which had only 80% power for models with greater than 0.051 heritability. The GPNN also outperformed the BPNN when applied to models containing functional and nonfunctional SNPs. Complex nonlinear interactions with binary endpoints that have previously been analyzed by logistic regression and classification and regression trees (CARTs) can be examined by GPNN. Motsinger et al. [30] demonstrated the use of grammatical evolution neural networks (GENNs) in detecting gene-gene and gene-environment interactions in high dimensional data with noise. GENN were found to be more vigorous with missing data and genotyping errors.

On the other hand, random forest (RF) algorithm is a flexible supervised machine learning algorithm that can be used for classification and regression. The RF algorithm is often able to produce good results even with missing values in the data and without need for hyper-parameter tuning. Therefore, RF algorithm can be well suited for high-dimensional genomic data analysis. This algorithm is also useful in reducing the search space of epistatic interactions, thereby creating a manageable set of possible combinations of genetic variants.

Finally, support vector machine (SVM) is a machine learning algorithm that uses hyper-planes for classification and regression tasks. The SVM approach has been applied to detecting gene-gene interactions through learning from the features of genetically interacting pairs. For training, SVM takes in two sets of feature vectors

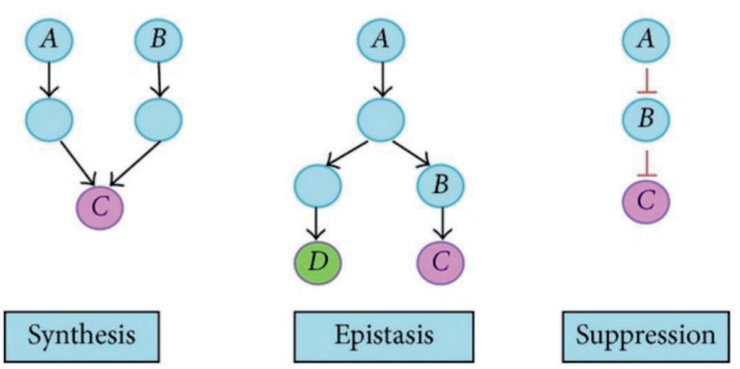

Figure 3.
Categories of gene-gene interactions retrieved from Koo et al. [23].

labeled as positive and negative, indicating presence and absence of genetic interaction, respectively. Feature mapping is done by use of a hyper-plane with maximum margin to separate genetically interacting pairs and non-genetically interacting pairs. SVM and neural network modeling was used to investigate gene-gene interactions in a study by Matchenko-Shimko and Dube [31]. They used pre-selection of SNP-SNP combination to determine the effects of interactions between genes. However, the pre-selection strategy did not work well with combinations of low disease allele frequencies and low margin effects. It was discovered that larger sample sizes are required for determining gene-gene interactions with SNPs having low marginal effect sizes as compared to interactions with moderate marginal gene effect sizes. Both SVM and ANN models exhibited good performance in increasing allele frequency with low marginal gene effects [31]. SVM was used to identify the most promising SNPs and interactions. Shen et al. used it in two stages for determining gene-gene interactions where the second stage involves the application of logistic regression analysis. It was shown that SVM is also useful in methods for case-control studies in which multiple logistic regression performs better than traditional logistic regression for each interaction. Additionally, application of the SVM in improving the accuracy of cancer classification, through extending the SVM pedigree-based generalized multifactor dimensionality, has been functional in detecting gene-gene and gene-covariate interactions in limited family samples [32]. Moreover, the SVM can also be used to extract known gene-disease associations and infer known genes for future experimental analysis using automatic literature mining based on dependency parsing and SVM [33].

In addition, the application of SVM in SUPPORTMIX [34], which is a local ancestry inference method, facilitates gene-gene and gene-environment interactions. For instance, Aschard et al. [35] highlighted that local ancestry estimates might provide insights into detecting gene-gene interactions, while Florez et al. [36] showed that non-European ancestry in the Latino populations is associated with type 2 diabetes and lower economic status, illustrating gene-environment interaction. Local ancestry inference estimates the proportion of alleles that originates from a particular population at every chromosomal site of an admixed individual. SUPPORTMIX integrates SVM with hidden Markov models (HMMs). Using SVM in SUPPORTMIX improves multi-way local ancestry inference overall, since it addresses the challenge of few genotyped or existing reference panels [1]. Furthermore, it facilitates both gene-gene and gene-environment interactions due to the improved computational time as a result of its flexibility and ability to handle "big data."

2.5 Elucidating disease-causing genetic variants

The identification of disease-causing genetic variants is challenging because several of them are found in the non-coding regions of the genome. The role of non-coding regions in the maintenance of genome functions is not well understood. However, some machine learning algorithms have been designed to annotate coding and non-coding genetic variants in order to identify disease-causing mutations. *Combined annotation-dependent depletion* (CADD) is an algorithm designed to annotate coding and non-coding variants [37]. CADD trains a linear kernel support vector machine to separate observed genetic variants from simulated ones. However, due to the SVM's inability to capture nonlinear relationships among features, a deep neural network that uses the same feature set and training data as CADD is preferred. Deep neural networks are better suited than SVMs for problems with large samples and features.

How genetic variants, especially those which are not within protein coding regions, affect RNA splicing is not entirely understood. This type of problem can

however be addressed by machine learning computational models designed to predict splicing during gene expression. Regulation of splicing is very important and faulty regulation could lead to several diseases, such as cancer and neurological disorders. A computational technique, that scores the magnitude of the effects of genetic variants on RNA splicing, was developed by Xiong et al. [38]. The computational model can be applied to any sequence with a triplet of exons and used to determine how splicing is altered by genetic variants. The model computes a score that predicts how much a given variant affects splicing.

Linkage and *association analysis* are types of neural network methods used to identify genes associated with diseases. Linkage analysis is used to detect the connection between a disease locus and a marker and uses genotypes as inputs and the outputs are phenotype values such as disease status and quantitative clinical variables. Association analysis on the other hand is used for detecting the disequilibrium between disease locus and marker. The data in association analysis are of case-control type with a sample comprised of genotypes for multiple markers. In most cases, it is useful to integrate genotype information into pathway analysis for more effective biological interpretation of these genotype contributions into the trait under consideration. In this case, *random survival forest pathway hunting* algorithm can be used to identify signaling pathways in a relatively small sample size [39].

Finally, considering the RF features, the RF algorithm can also be used in identifying a set of risk-associated SNPs from a large number of unassociated SNPs in models of complex diseases. There are unknown interactions among true risk-associated SNPs or SNPs and the environment in large-scale genetic data and RF can be used to significantly reduce the number of SNPs in the data as pointed out previously.

2.6 Applying learning algorithms in clinical decision process

Setting appropriate diagnostic and effective therapeutic regimens is a critical clinical decision and essential for setting effective health measures and efficient strategies to control a disease. This process is limited by the lack of advanced diagnostic tools and approved therapy or vaccine against most existing and emerging diseases [40, 41]. Moreover, despite undeniable advances made in understanding of human biology, etiology, and pathogenesis of several diseases, and emergence of advanced technologies, the translation of the existing biological knowledge toward effective new treatments and clinical interventions has not been as fast as expected or anticipated. This highlights the need for powerful and general tools for orienting these clinical decision processes. Machine learning algorithms are contributing to satisfying this need with several advantages in representational power even though challenges in biological interpretation still hamper clinical applications [15].

As an initial illustration, Adabor and Acquaah-Mensah [42] introduced the median supplement model to appropriately balance a training set with unequal numbers of instances associated with each class or group to improve the classification decision. They also assessed different machine learning techniques in predicting the receptor expression status of breast cancer patients, namely progesterone receptor (PR) status and HER2 expression status using gene expression datasets. These receptors are essential in deciding on treatment and predicting the treatment outcome. In this chapter, we used results of their performance evaluations to highlight two essential features common to most of the machine learning algorithms as shown in **Figure 4**: (1) as the size of the training set increases, the performance of the learning algorithm increases (see Sample Data 1 vs. Sample Data 2) and (2) learning algorithm on a balanced training set may perform better than on an unbalanced training set (see NB vs. MNB and RF vs. MRF).

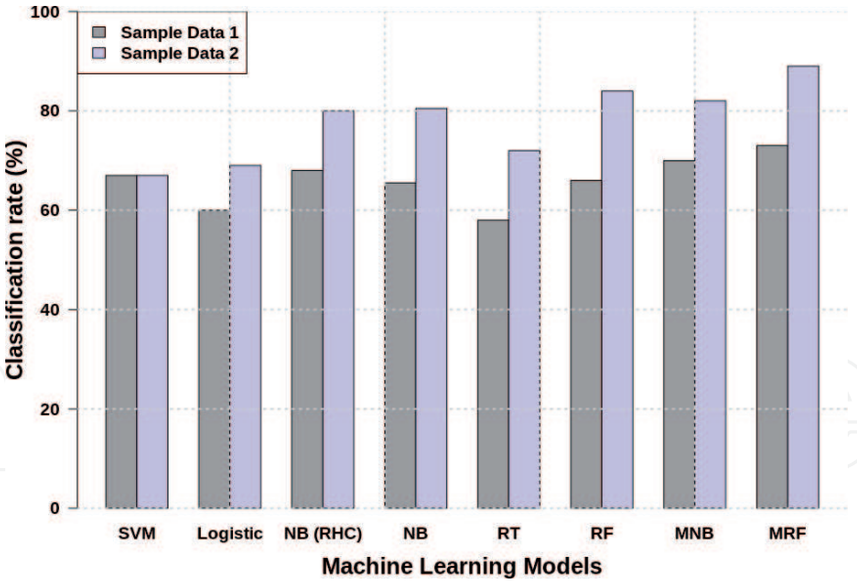

Figure 4.
Performance of different machine learning techniques for predicting progesterone receptor (PR) status phenotype of breast cancer patients based on classification rate (proportion of correctly classified instances), information extracted from [42]. Sample Data 1 is a smaller-sized dataset as compared to sample data 2, containing 162 and 1146 instances of breast cancer patients, respectively. Learning techniques: support vector machine (SVM), logistic regression (logistic), Bayesian network (BN), Naive Bayes (NB), random trees (RT), random forest (RF), median-supplement Naive Bayes (MNB), and median-supplement random forest (MRF).

It is worth mentioning that machine learning algorithms have been used in several contexts with a common goal of improving healthcare measures and patient clinical management. For examples, deep learning algorithms are used to classify patients based on clinical healthcare records [43], to predict the effectiveness of clinical trials (i.e., likelihood of success or failure of clinical trials) [44], to improve and predict patient treatment response and outcome based on pharmaco-genomics data [45]. Moreover, Nemati et al. [14] optimized a treatment dosing policy for intensive care patients using deep reinforcement learning and Wang et al. [46] predicted drug-target binding site interactions using ANN with two hidden layers taking a drug and a target binding site as inputs. Finally, it is known that drug repositioning or re-purposing approach, which examines new therapeutic uses for approved drugs, represents an optimal model for suggesting new drugs using drug-target interactions [40, 41]. Wang and Zeng [47] used a learning technique based on restricted Boltzmann machines to predict novel drug-target interactions directing to drug re-purposing.

3. Integrative approaches for post-genomic analysis

Over the years, thousands of genetic associations have been discovered using genetic approach, known as *genome-wide association studies* (GWAS). GWAS approaches are mostly based on a single-marker association test model that leverages thousands of genomes of cases and controls (sick and healthy individuals) in order to elucidate variants or single-nucleotide polymorphisms (SNPs) with unusual significant differences in frequency throughout genomes [48]. This indicates that GWAS approaches are based on machine learning techniques, which mostly take SNP profiles of cases and controls as inputs, and predict a SNP carrying disease risk. Note that these approaches have been successful [49] and several GWAS results have helped elucidating genetic determinants of susceptibility to several diseases, including complex diseases, such as cancer, and monogenic diseases, such as sickle cell

disease. In fact, in case of the breast cancer disease, a genetic testing tool has been implemented [50] based on specific genetic variants in breast cancer type 1 (BRCA1) and 2 (BRCA2) susceptibility genes in chromosomes 17 (17q21.31) and 13 (13q13.1) [51], respectively. It is widely known that the outcome of a disease, in particular a complex disease, or a response to a drug is influenced by multiple genes and significant contribution from the environment. This strongly argues that using only genomic analysis will not be sufficient to entirely embed phenotypic variation and heritability, suggesting that genomic analysis alone is not sufficient to elucidate the complex structure of the disease [52]. Thus, there is a significant need of integrating information derived from environmental studies and other heterogeneous datasets into genomic analysis to enhance the predictive power of genomic analysis.

As indicated above, even though genomic information is critical, it is not sufficient to completely elucidate disease outcome and progression, which involve gene-gene and gene-environment interactions. In this context, the post-genomic analysis may provide a new paradigm to genomic analysis and may enable further functional characterization of genetic susceptibility to a disease and correlate disease-associated (candidate) genes by combining association signals from genomic analysis and available knowledge, including functional, environmental, epidemiological, and clinical information. This integrative approach increases the likelihood of effectively identifying suitable candidate genes [53] and biological pathways that may be critical in the etiology and pathogenesis of the disease, and in the drug response. The next goal is to integrate large-scale datasets from heterogeneous sources [2, 54] to move beyond a single genomic approach and foster a whole genome-based integrative approach to achieve global view [55]. A biological network, which is a network modeling a biological system as an entity composed of sub-units connected as a whole, has become a useful tool enabling the integration of heterogeneous datasets into a single framework [26].

4. Challenges and perspectives

Currently, there is an exponential growth of several platforms producing large-scale datasets, including genomics, transcriptomics, proteomics, metabolomics, microbial and epidemiological data. These high-dimensional datasets from heterogeneous sources create an opportunity of designing appropriate data-driven learning algorithms and models to ensure effective post-genomic medicine and biomedical research with an increased prediction power. While the use of these large-scale post-genomic datasets from heterogeneous sources, such as transcriptomics, proteomics, metabolomics, microbial and epidemiological data, shows several potential advantages and opportunities, many challenges still exist in terms of computational models, learning algorithms, and biological interpretation of result outputs. Furthermore, as discussed previously, learning, reinforcement, and deep learning algorithms are quickly evolving with several potential applications in biology and medicine (see **Section 2.6**). Currently, predictions from different models are unable to contribute to clinical decision processes as the effectiveness of these models still poses problems in the absence of ground-truth, gold standard (benchmark) datasets, or experimental validation. This suggests that one of the future trend aspects of learning algorithms in biology and medicine will be to make possible the integration of predictive models generated by these learning algorithms into dynamic clinical settings. This integration will necessitate that issues raised above are addressed systematically and will ensure an effective exploitation of the post-genomic datasets and potentially revolutionize the study of human disease and health.

Machine intelligence and deep learning models present more powerful computational techniques that are able to effectively learn from large complex datasets in order to reveal several hidden interactions within cell variables and give more insight into the intricate processes linked to diseases [56]. On the other hand, despite the current undoubted data wealth, we still have a very limited understanding of the mechanisms underlying the outcome, pathogenesis, and progress of many diseases, which is reflected in an existing gap between this data wealth and translation toward enhancing treatment and interventions for diseases, leading to the paradigm of "world with data wealth and information poor". This is partly due to issues related to different existing datasets, including: (1) increased heterogeneity within a dataset as, in general, these datasets are collected across different locations, thus lacking a standardized representation of the data and (2) variation of cohorts in terms of size across populations and geographical locations. This highlights the need for designing adequate meta-analysis models to assist in retrieving useful information within each data source. This may also require more advanced machine learning techniques to play an important role in genomic medicine and advance our knowledge about disease and health.

5. Conclusions

Numerous large-scale platforms have been designed for producing different types of high-dimensional datasets, including genomics, transcriptomics, proteomics, metabolomics, microbial and epidemiological data. This data deluge provides a rich source of information, which can advance our understanding of human and pathogenic organisms to enhance post-genomic medicine and biomedical research. In this chapter, we have provided some illustrations of machine learning algorithms for knowledge discovery in biological and health areas and discussed existing challenges. This discussion highlights the need for adequate meta-analysis-based post-genomic models to optimally integrate diverse datasets from different sources. This clearly suggests that initial machine learning algorithms will need to be refined or new ones need to be developed to account for current data challenges in order to speed up the translation of the current and future knowledge into effective new treatment strategies and health measures, enabling efficient clinical disease management and ensuring effective post-genomic medicine.

Conflict of interest

The authors declare that they have no competing interests.

Author details

Gaston K. Mazandu[1,2,3*], Irene Kyomugisha[2,4], Ephifania Geza[2,3], Milaine Seuneu[2], Bubacarr Bah[2,4] and Emile R. Chimusa[1]

1 Division of Human Genetics, Department of Pathology, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town (UCT), Cape Town, South Africa

2 African Institute for Mathematical Sciences (AIMS), Cape Town, South Africa

3 Computational Biology (CBIO) Division, Department of Integrative Biomedical Science, Institute of Infectious Disease and Molecular Medicine, Faculty of Health Sciences, University of Cape Town (UCT), Cape Town, South Africa

4 Division of Applied Mathematics, Department of Mathematical Sciences, University of Stellenbosch, Stellenbosch, South Africa

*Address all correspondence to: kuzamunu@aims.ac.za

References

[1] Geza E, Mugo J, Mulder NJ, Wonkam A, Chimusa ER, Mazandu GK. A comprehensive survey of models for dissecting local ancestry deconvolution in human genome. Briefings in Bioinformatics. 2018:1-16

[2] Mazandu GK, Chimusa ER, Mulder NJ. Gene ontology semantic similarity tools: Survey on features and challenges for biological knowledge discovery. Briefings in Bioinformatics. 2016;**18**(5):886-901

[3] Strobl C, Malley J, Gerhard T. Characteristics of classification and regression trees, bagging and random forests. Psychological Methods. 2009;**14**(4):323-348

[4] Altman NS. An introduction to kernel and nearest-neighbor nonparametric regression. The American Statistician. 1992;**46**(3):175-185

[5] Khondoker M, Dobson R, Skirrow C, Simmons A, Stahl D. A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. Statistical Methods in Medical Research. 2016;**25**(5):1804-1823

[6] Cortes C, Vapnik V. Support-vector networks. Machine Learning. 1995;**20**(3):273-297

[7] Breiman L. Random forests. Machine Learning. 2001;**45**(1):5-32

[8] Karpathy A. CS231n Convolutional Neural Networks for Visual Recognition. Available from: http://cs231n.github.io/neural-networks-1/#nn 2017

[9] Murtagh F, Contreras P. Algorithms for hierarchical clustering: An overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. 2012;**2**:86-97. 7 (2017) e1219

[10] Hartigan JA, Wong MA. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society, Series C. 1979;**28**(1):100-108

[11] Bothe MK, Dickens L, Reichel K, Tellmann A, Ellger B, Westphal M, et al. The use of reinforcement learning algorithms to meet the challenges of an artificial pancreas. Expert Review of Medical Devices. 2013;**10**(5):661-673

[12] Weng WH, Gao M, He Z, Yan S, Szolovits P. Representation and reinforcement learning for personalized glycemic control in septic patients. In: 31st Conference on Neural Information Processing Systems (NIPS); Long Beach, CA, USA. 2017

[13] Ling Y, Hasan SA, Datla V, Qadir A, Lee K, Liu J, et al. Learning to diagnose: assimilating clinical narratives using deep reinforcement learning. In: Proceedings of the 8th International Joint Conference on Natural Language Processing. 2017. pp. 895-905

[14] Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In: Conference Proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. 2016. pp. 2978-2981. DOI: 10.1109/EMBC.2016.7591355

[15] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, et al. Opportunities and obstacles for deep learning in biology and medicine. Journal of the Royal Society Interface. 2018;**15**. DOI: 10.1098/rsif.2017.0387

[16] Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nature Biotechnology. 2015;**33**(8):831-838

[17] Urda D, Montes-Torres J, Moreno F, Franco L, Jerez JM. Deep learning to analyze RNA-seq gene expression data. In: International Work-Conference on Artificial Neural Networks. Springer; 2017. pp. 50-59

[18] Dincer AB, Celik S, Hiranuma N, Lee S. DeepProfile: Deep learning of cancer molecular profiles for precision medicine. bioRxiv 2018. 278739. DOI: 10.1101/278739

[19] Wang L, Xi Y, Sung S, Qiao H. RNA-seq assistant: machine learning based methods to identify more transcriptional regulated genes. BMC Genomics. 2018;**19**:546

[20] Rosario SF, Thangadurai K. RELIEF: Feature selection approach. International Journal of Innovation Science and Research. 2015;**4**(11):218-224

[21] Mazandu GK, Mulder NJ. Generation and analysis of large-scale data-driven Mycobacterium tuberculosis functional networks for drug target identification. Advances in Bioinformatics. 2011;**2011**:801478

[22] Rapanoel HA, Mazandu GK, Mulder NJ. Predicting and analyzing interactions between Mycobacterium tuberculosis and its human host. PLoS One. 2013;**8**(7):e67472

[23] Ahmed I, Witbooi P, Christoffels A. Prediction of human-*Bacillus anthracis* protein-protein interactions using multi-layer neural network. Bioinformatics. 2018;**34**(24):4159-4164

[24] Pavlopoulos GA, Secrier M, Moschopoulos CN, Soldatos TG, Kossida S, Aerts J, et al. Using graph theory to analyze biological networks. BioData Mining. 2011;**4**:10

[25] Alm E, Arkin PA. Biological networks. Current Opinion in Structural Biology. 2013;**13**:193-202

[26] Ma'ayan A. Introduction to Network Analysis in Systems Biology. Science Signaling. 2011;**4**(190):tr5

[27] Mazandu GK, Mulder NJ. Enhancing drug target identification in Mycobacterium tuberculosis. In: Tuberculosis: Risk Factors, Drug Resistance and Treatment. NOVA Publishers; 2012

[28] Koo CL, Liew MJ, Mohamad MS, Salleh AHM. A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology. BioMed Research International. 2013;**2013**:432375

[29] Ritchie MD, White BC, Parker JS, Hahn LW, Moore JH. Optimization of neural network architecture using genetic programming improves detection and modelling of gene-gene interactions in studies of human diseases. BMC Bioinformatics. 2003;**4**:28

[30] Motsinger-Reif AA, Fanelli TJ, Davis AC, Ritchie MD. Power of grammatical evolution neural networks to detect gene-gene interactions in the presence of error. BMC Research Notes. 2008;**1**:65

[31] Matchenko-Shimko N, Dube MP. Gene-gene interaction tests using SVM and neural network modeling. In: 2007 Proceedings of the IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology. 2006. pp. 90-97

[32] Fang Y, Chiu Y. SVM-based generalized multifactor dimensionality reduction approaches for detecting gene-gene interactions in family studies. Genetic Epidemiology. 2012;**36**(2):88-98

[33] Ozgur A, Vu T, Erkan G, Radev DR. Identifying gene-disease associations using centrality on a literature mined gene-interaction network. Bioinformatics. 2008;**24**(13):i277-i285

[34] Omberg L, Salit J, Hackett N, Fuller J, Matthew R, Chouchane L, et al. Inferring genome-wide patterns of admixture in Qataris using fifty-five ancestral populations. BMC Genetics. 2012, 2012;**13**(1):49

[35] Aschard H, Gusev A, Brown R, Pasaniuc B. Leveraging local ancestry to detect gene-gene interactions in genome-wide data. BMC Genetics. 2015;**16**:124

[36] Florez JC, Price AL, Campbell D, Riba L, Parra MV, Yu F, et al. Strong association of socioeconomic status with genetic ancestry in Latinos: Implications for admixture studies of type 2 diabetes. Diabetologia. 2009;**52**(8):1528-1536

[37] Quang D, Chen Y, Xie X. DANN: A deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics. 2014;**31**(5):761-763

[38] Xiong HY, Alipanahi B, Lee LJ, Bretschneider H, Merico D, Yuen RKC, et al. The human splicing code reveals new insights into the genetic determinants of disease. Science. 2015;**347**(6218):1254806

[39] Chen X, Ishwaran H. Pathway hunting by random forests. Bioinformatics. 2013;**29**(1):99-105

[40] Mazandu GK, Chimusa ER, Rutherford K, Zekeng EG, Gebremariam ZZ, Onifade MY, et al. Large-scale data-driven integrative framework for extracting essential targets and processes from disease-associated gene data sets. Briefings in Bioinformatics. 2018;**19**(6):1141-1152. DOI: 10.1093/bib/bbx052

[41] Rutherford KD, Mazandu GK, Mulder NJ. A systems-level analysis of drug-target-disease associations for drug repositioning. Briefings in Functional Genomics. 2017;**17**(1):34-41

[42] Adabor ES, Acquaah-Mensah GK. Machine learning approaches to decipher hormone and HER2 receptor status phenotypes in breast cancer. Briefings in Bioinformatics. 2017. DOI: 10.1093/bib/bbx138

[43] Huddar V, Desiraju BK, Rajan V, Bhattacharya S, Roy S, Reddy CK. Predicting complications in critical care using heterogeneous clinical data. IEEE Access. 2016;**4**:7988-8001. DOI: 10.1109/access.2016.2618775

[44] Artemov AV, Putin E, Vanhaelen Q , Aliper A, Ozerov IV, Zhavoronkov A. Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes. bioRxiv. 2016. doi: 10.1101/095653

[45] Kalinin AA, Higgins GA, Reamaroon N, Reza SM, Allyn-Feuer A, Dinov ID, Najarian K, Athey BD. Deep learning in pharmacogenomics: From gene regulation to patient stratification. 2018. https://arxiv.org/abs/1801.08570v1

[46] Wang C, Liu J, Luo F, Tan Y, Deng Z, Hu QN. Pairwise input neural network for target-ligand interaction prediction. In: 2014 IEEE International Conference on BIBM. 2014. pp. 67-70

[47] Wang Y, Zeng J. Predicting drug-target interactions using restricted Boltzmann machines. Bioinformatics. 2013;**29**:i126-i134

[48] Chimusa ER, Dalvie S, Dandara C, Wonkam A, Mazandu GK. Post genome-wide association analysis: Dissecting computational pathway/network-based approaches. Briefings in Bioinformatics. DOI: 10.1093/bib/bby035

[49] Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Research;**42**:D1001-D1006

[50] Gabai-Kapara E, Lahad A, Kaufman B, Friedman E, Segev S, Renbaum P,

et al. Population-based screening for breast and ovarian cancer risk due to BRCA1 and BRCA2. PNAS. 2014;**111**(39):14205-14210

[51] Försti A, Luo L, Vorechovsky I, Söderberg M, Lichtenstein P, Hemminki K. Allelic imbalance on chromosomes 13 and 17 and mutation analysis of BRCA1 and BRCA2 genes in monozygotic twins concordant for breast cancer. Carcinogenesis. 2001;**22**(1):27-33

[52] Chimusa ER, Mbiyavanga M, Mazandu GK, Mulder NJ. AncGWAS: A post genome-wide association study method for interaction, pathway, and ancestry analysis in homogeneous and admixed populations. Bioinformatics. 2016;**32**(4):549-556

[53] Ma X, Gao L. Biological network analysis: Insights into structure and functions. Briefings in Functional Genomics. 2012;**11**(6):434-442

[54] Mulder NJ, Akinola RO, Mazandu GK, Rapanoel H. Using biological networks to improve our understanding of infectious diseases. Computational and Structural Biotechnology Journal. 2014;**11**(18):1-10

[55] Mazandu GK, Opap K, Mulder NJ. Contribution of microarray data to the advancement of knowledge on the Mycobacterium tuberculosis interactome: Use of the random partial least squares approach. Infection, Genetics and Evolution. 2011;**11**(4):725-733

[56] Leung MKK, Xiong HY, Lee LJ, Frey BJ. Deep learning of the tissue-regulated splicing code. Bioinformatics. 2014;**30**(12):i121-i129