Available online at www.worldnewsnaturalsciences.com

World News of Natural Sciences

An International Scientific Journal

WNOFNS 44 (2022) 43-62 EISSN 2543-5426

Applications and Limitations of Integrative Robust Approaches in Multiple Omics Analysis

Jesujoba Owolabi[1,*], Yagoub Adam[2], Titilope Dokunmu[3], Ezekiel Adebiyi[4], Nwankwo Chukwuma[5]

[1]Covenant Applied Informatics and Communication African Centre of Excellence,
Department of Computer and Information Sciences, Covenant University, Ota, Ogun State, Nigeria
*E-mail address: Jesujoba.owolabi.ace@stu.cu.edu.ng

[2]Covenant University Bioinformatics Research (CUBRe), Covenant University,
Ota, Ogun State, Nigeria
E-mail address: yagoub.adam@covenantuniversity.edu.ng

[3]Covenant Applied Informatics and Communication African Centre of Excellence,
Department of Biochemistry, Covenant University, Ota, Ogun State, Nigeria
E-mail address: titilope.dokunmu@covenantuniversity.edu.ng

[4]Covenant Applied Informatics and Communication African Centre of Excellence,
Department of Computer and Information Sciences, Covenant University, Ota, Ogun State, Nigeria
E-mail address: Ezekiel.adebiyi@covenantuniversity.edu.ng

[5]Department of Computer and Information Sciences, Covenant University, Ota, Ogun State, Nigeria
E-mail address: chumasalem@gmail.com

ABSTRACT

Comprehensive biological research shows that genomic information garnered over the years is not enough to completely understand biological systems even at the cellular level. Integrative omics focuses on the integration of multiple omics data types, with an unceasing improvement of high-content, real-time, multimodal, multi-omics technologies. This will lead to a deep understanding of biological systems. Multi-omics can be used to profile genetic, transcriptomic, epigenetic, spatial, proteomic and lineage information in single cells. This transformative method provides bioinformatics and integrative methods that can be used through multiple types of data, and it can identify relationships within cellular modalities, provide a deeper representation of cell state, and aid assembly of data sets to provide useful knowledge. Here, we discuss the challenges of multiple omics datatype integration, limitations of the complex machine learning models and recent technology advances in multi-omics data integration.

(Received 16 June 2022; Accepted 08 July 2022; Date of Publication 09 July 2022)

Keywords**:** Multi-omics, Biological-systems, Integration, Machine Learning

1. INTRODUCTION

Multi-omics is a new biological analysis approach where multiple "omes" datasets are integrated into one set of “omes”. Such multiple “omes” include the genome, transcriptome, proteome, metabolomes, epigenomes and microbiome [1]. The study of DNA, mRNA and proteins which makes up the central dogma of a living organism, can be broadly denoted as genomics, transcriptomics and proteomics. Genomics which studies the genetic blueprint of a cell entails investigating the DNA to identify the presence or absence of certain genes [2].

Genomics approaches have been extensively explored in identifying the genes and genetic loci involved in the development of human diseases [3] while transcriptomics studies the transcribed genetic material i.e. all RNA transcripts including ribosomal RNA, messenger RNA and other non-coding RNA, which are actively expressed during certain processes to broadly reflect the functional state of the cell [2].

Proteomics consists of all the proteins that are expressed in a cell and they help in understanding the flow of information within the cell. In the central dogma process, the matured RNA is transcribed into proteins. The structure of a protein is quite complex, they exhibit different conformational structures, different interactions and localizations depending on temporal and spatial factors. Other aspects include lipidomics (which study the complex analysis of lipids) and glycomics; the study of all glycan structures of an organism, although they are not a part of the central dogma analysis [2]. The constant advancement of high throughput technologies in generating diverse types of omics data has contributed tremendously to human health and the treatment of diseases [4].

Although extensive simultaneous analysis has been conducted at different omics levels including gene expression, mRNA, copy number variations, microRNA and DNA methylation [5], integrative omics analysis offers a more holistic and powerful way of improving the strength of single omics analysis efficiently enough to study life in a concerted way [6].

Integrative omics analysis focuses on the integration of multiple omics data types for the same cohort of samples [6], For example, analysis of multiple datasets can be used for detecting suicidal markers in depressed patients. In other words, the single omics analysis can be referred to as horizontal integration (the integration of a specific type of omics data across different samples.

Biomedical data are becoming increasingly complex, from the quantity to the quality of the data. High-throughput data acquisition techniques and their digitization have massively increased sample number, whereas heterogeneity includes biologically related features from clinical metadata and multiple omics data type. Furthermore, data could be collected via multiple platforms [7], thereby introducing bias, complexity, and noise. Some other omics-specific problems include ethical standards, study framework, management of data sharing, reproducibility of the research [2] etc.

Machine learning methods are therefore appropriate for data modelling and integration of multi-omics data in these situations [3]. However, it is important to apply algorithms that suit the data to be used whether it’s a classification or clustering algorithm [8].

2. CHALLENGES OF MULTI OMICS

2. 1. Curse of dimensionality

Current trends have identified integrative analysis of multi-omics as the next step in understanding complex biological systems. The inclusion of different omics data would bring about new information in the functions of pathways and systems associated with a disease trait, although this process is not void of challenges. Dimensionality reduction methods are frequently applied in omics studies because datasets from each level are faced with the q >> n problem, where the number of features in a study increases with no proportional increase with the samples [9]. This curse of dimensionality problem makes most robust algorithms vulnerable to the overfitting problem [10]. Random noise in samples may be prone to overfitting issue and lead to poor generalization performance [5]. Dimensionality reduction-based methods in machine learning are used to perform feature selection or feature extraction [11]. Feature extraction transforms datasets from a higher dimensionality to a lower dimensionality, whereas selection of features works by reducing dimensionality through identifying only a group of important features [9], [12], [13]. Therefore, in employing dimension reduction techniques, important features that can be used for prediction and model performance is improved as multi-collinearity between features is removed.

2. 2. Data heterogeneity and Data Missingness

Another challenge associated with the incorporation of diverse types of data in a single model is heterogeneity. One of many reasons for data heterogeneity is the variability of platforms used in generating multi-omics data and the different data storage and formats. The majority of multi-omics integrative analysis tools require data in specific formats, most notably the Feature X Sample matrix [14], therefore the problem of data missingness in the construction of a model reduces the model's predictive power and can lead to spurious relationship or correlations. Removal of missing observations is a common way to handle missing data, but it is very expensive when patient samples are already very limited. Data imputation based on known data is a better way to solve this problem. Others include the difference in attributes, scaling and distribution, and a variety of multi-modality in the data such as the undirected categories, intervals [9]. Therefore, individual omics data need to be pre-processed, such pre-processing process includes data filtering, normalization (this can handle mismatched distribution), quality control checks etc. before integration[14]. Note that without proper normalization, additional weights would be given to more features and there would be high noise [15].

2. 3. Class imbalance

ML-based models in omics studies are frequently challenged the class imbalance problem [9], [16]. For instance, a machine learning classifier trained to predict the location of genes causing antibiotic resistance in the genetic makeup of the organism may suffer from the class inequality problem, which means that the data to be used contained more negative or control samples than case or positive samples. Take, for instance, the protein structures dataset used in predicting interactions between amino acid pairs can suffer from rarity or imbalance problem due to the sparseness of the contacts [17]. Other omics prediction-based problems includes: Post translation modifications [18], Protein-DNA binding residue [19], DNA methylation sites [20], [21], protein-protein interactions [22], [23] and functional antimicrobial peptides [24] etc.

2. 4. Scalability issues

ML algorithms can create models whose performance improves as more data becomes available. However, large data types collected from many high-throughput omics systems may pose scalability issues. Executing multiple omics analysis based on Machine Learning methods on a single computer is becoming increasingly difficult. However, advancements in big data optimization algorithms, real-time ML, parallel processing in ML algorithms, and large-scale cloud computing analysis is now possible [9].

3. ROBUST APPROACHES IN SOLVING MULTI-OMICS INTEGRATION CHALLENGES

Studies have established the importance of omics data type integration in revealing which biological pathways vary between the target and the control group, therefore the analysis of only a genomic or proteomic data type would limit these correlations and provide a partial view of the complex biological system. By combining various types of data, researchers can overcome the limitations of individual studies and better identify disease-causing variants and their downstream molecular targets. However, the combined examination of multi-omics data is not void of computational challenges (problems already discussed above), it intensifies the challenges linked to single-cell omics. Therefore, robust models are needed to indeed execute integrative analysis efficiently. In this regard, machine learning-based approaches have been employed as a key player in circumventing the specific computational challenges involving multiple omics data [25], [26].

3. 1. Dimension reduction-based methods

Dimension reduction methods assume that the data has an inherent low dimensional representation, with the low dimension frequently corresponding to the number of variables. Machine learning-based Dimension reduction techniques may be categorized into feature extraction (FE) and feature selection (FS). Feature selection selects one of the smallest sets of features guaranteeing the highest classification performances. Alternatively, the maximal set including all the relevant features can be chosen, this is known as feature extraction. Variable selection has been widely used to analyze single-level omics data, where the dimensionality of omics features is typically much greater than the sample size. Identifying a subset of important features usually results in improved interpretability and improved prediction using the chosen model. Both are necessary for the success of the combined analysis of multi-omics data. This explains, at least in part, why the variable selection is one of the most powerful and widely used data integration tools [6].

Principal component analysis (PCA), joint non-negative matrix factorization (NMF), multi-omics factor analysis (MOFA) and multiple co-inertia analysis (MCIA) are exemplars of machine learning feature extraction models that are used in the integrative analysis. These FE methods can capture linear interactions in the data. Nonlinear relationship methods including representation learning, t-SNE and autoencoder exist etc. ML-based feature selection techniques are broadly categorized into filter, wrapper and embedded methods. Filter methods like correlation-based FS (CFS), maximal-relevance and minimal-redundancy (mRMR), ReliefF [27] and Information Gain are employed as a pre-processing step before training any model, while wrapper methods such as support vector machine recursive feature elimination (SVM-RFE) [28] and Boruta [29] can automatically judge the importance of features, Boruta can work so well without prior specific input by the user to extract importance. Embedded methods that comprise the least absolute shrinkage and selection operator (LASSO) [30], Elastic Net [31], stability selection, etc. can carry out feature selection, a portion of the model building process. Extraction of Features is commonly used in the unsupervised integrative analysis, i.e. when the target labels are not known [9]. In multi-omics studies, FE can aid in the identification of disease subgroups. Many feature extraction methods for integrative omics exploratory analysis have been proposed in recent years, with many of them based on PCA [32]. A specific advantage of using a linear model is that they give some interpretations for the features in each group that is more observed, unlike some other methods, for instance, the similarity-based dimension technique, which ignores the original feature if the similarity between the sample is scored [11].

3. 2. Data Heterogeneity

As earlier stated, the integration of different data types into a single model for prediction is one of the biggest challenges due to the heterogeneity of the data. Previous studies have outlined the application of machine learning algorithms in handling heterogeneous data in many ways. These algorithms include Penalized linear models such as LASSO, ElasticNet and TANDEM, Decision trees and Random Forest for tree-based models, Bayesian multitask and simple multitask as Multiple kernel learning models, Graphs and Networks (SNF, NetlCS, PARADIGM, HetroMed), Latent Sub-space clustering (iCluster+, Scluster, MV-RBM) and Deep learning (Multimodal DBN, Multimodal DNN, Improved CPR, AuDNNsynergy). Other new models designed in investigating multiomics includes, iClusterBayes [33], Bayesian multiple kernel learning (BMKL) [34] was used to integrate data from different profiling sources (CNA, DNA methylation, gene expression, reverse phase protein array (RPPA)) for the prediction of drug sensitivity in breast cancer cell lines.

Table 1. Machine learning models in addressing multi-omics heterogeneity issue.

APPROACH	ALGORITHMS	DESCRIPTION	USE-CASE	REFERENCES
Network-based	Similarity network fusion (SNF)	It works by combining individual similarity from various data sources or types to create a single network that captures the complementary information.	Another study, incorporated SNF, JIVE, MCIA, MFA and MCCA to investigate multi-omics data. It suggested that multi-omics data integration largely benefits from a feature selection step and that SNF is a robust method than others.	[35]

	Network-based integration of multi-omics data (NetICS)	Integrates multi-omics data on a directed functional interaction network. It uses a per-sample network diffusion model on a directed functional interaction network and derives a population-level gene ranking by aggregating individual rankings and provides a global ranking for all samples	The heterogenous multi-omics data was integrated into a directed interaction network in the prioritization of cancer genes	[36]
	Pathway Recognition Algorithm using Data Integration on Genomic models (PARADIGM)	Probabilistic graphical models of cellular pathways by using a factor graph to represent the relations between the entities within the pathways	PARADIGM approach was used for analyzing gene expression and copy number data from TCGA Glioblastoma (GBM) revealing 4 subtypes of the disease. Another study showed the application of PARADIGM to derive novel insights into breast cancer using copy number and gene expression data	[37]
	Heterogenous-information networks (HetroMed)	Heterogenous information networks can handle any kind of data and it's mostly used for medical diagnoses.	The model was used to extract latent low dimensional features from clinical record data for robust medical diagnosis.	[38]
Concatenation-based Multiple Kernel learning	Simple multiple kernel learning (simple MKL)	The simple and Bayesian multiple kernels employ individual kernel function on different sources of data.	Simple MKL algorithm was applied to different data types in detecting glioblastoma multiforme.	[39], [40]
Penalized-Linear based	ElasticNet and Least absolute shrinkage	TANDEM, an ElasticNet based two-stage model is useful	Elastic Net was applied to investigate drug-response	[41], [42]

	and selection operator (LASSO), TANDEM	when data sources with infinite attributes outnumber data sources with binary attributes.	obtained from the CCLE using multi-omics data TANDEM uses a two-stage feature selection approach, with the first stage utilizing all binary variables, referred to as upstream data, and the second stage utilizing continuous gene expression variables referred to as downstream data.	[43]
Latent Sub-space Clustering-based	iCluster	Based on joint latent variable modelling or integrative clustering. Data originate from low dimensional representation, which determines the distribution of the observed data	iCluster was applied to two cancer datasets breast cancer and lung cancer to identify clinically relevant disease subtypes in latent sub-space. 33 cancers in the pan-cancer analysis were obtained with over ten thousand tumours and iCluster was applied to cluster them. iCluster used in the identification of cancer genes from the glioblastoma dataset, the clustering analysis showed three distinct subtypes.	[44] [45] [46]
	iCluster+	Cluster was upgraded to iCluster+ to include different data models for numeric, categorical and binary values i.e., it assumes different distribution for diverse data types	iCluster+ identified 12 distinct clusters using mutation, copy number, and gene expression profiles from 729 cancer cell lines representing 23 tumour types from CCLE.	[47]
	Mixed variable restricted Boltzmann machine (MV-RBM)	To cluster, the data are first transformed into latent sub-space and then clustering	In diabetes mellitus studies, data from highly heterogeneous sources such as	[48]

		analysis would be performed on the latent profiles	demographics, diagnosis, pathologies, and treatments were transformed into latent profiles (homogenous representation) using MV-RBM.	
Deep Learning	Deep Belief Network (DBN) Deep Neural Network (DNN) Improved Clustering and Page Rank (CPR) Autoencoders (AuDNNsynergy)	Deep learning is composed of multiple layers of artificial neurons, each with its weight value that is updated by the gradient descent algorithm during backpropagation to minimize the global loss function. DBN is mostly used in feature representation as it's built with multiple Boltzmann machines concatenated in a stacked manner with an input visible layer and an adjacent hidden layer trained with the aim to learn a probability distribution in the input set.	Multiple sources of omics data were combined with clinical data to perform integrated clustering using DBN. CPR was applied to multiple-omics data i.e., gene expression, DNA methylation, copy number, and somatic mutation data for five cancer types. Autoencoder was used in identifying the prognostic subtypes of high-risk neuroblastoma using CAN and gene expression data.	[49] [50] [51]

3. 3. Class Imbalance Learning models (CIL)

As earlier discussed, class imbalance problem arises as a result of disproportionality between the classes. Here, the minority or positive class which are the target always have smaller samples compared to the negative or majority class.

In other words, the proportionality in the two classes can be regarded as bias thereby favoring the majority class when the performance evaluation is measured. Therefore, robust models are needed in overcoming this challenge especially in the area of integrative analysis where multiple data from different sources are analyzed.

Class imbalance learning models can be grouped into data sampling and dimension reduction methods, ensemble modelling and cost-sensitive learning and asymmetric classification [52]. Data sampling method is the easiest way in tackling the class disproportionality problem. Data sampling can be categorized into the under-sampling of the negative class and the oversampling of the target class. Prior to applying the classifier, the dataset is sampled by randomly oversampling and random under-sampling or heuristically using one-sided selection. Random oversampling involves randomly oversampling the minority class by duplication, leading to the generation of new sample from current sample. This new samples obtained from the current sample can be created by an oversampling technique known as SMOTE (Synthetic Minority Oversampling Technique) [9] and CBO (Cluster based Oversampling) [53]. Although, data sampling is easy, it cannot be accounted for as a valid solution to the class imbalance problem because so many samples are forfeited. The one-sided selection (OSS) is an under-sampling technique that chooses to carefully exclude samples from the negative class i.e., samples that are correctly classified after random sampling and k-nearest neighbor while leaving out the minority class which is the target untampered. Another instance is the use of focal loss that penalizes the majority sample during loss calculation and give more weight to minority class. Dimension reduction techniques have been previously discussed and some have been engineered in the imbalance challenge. However, with the consistent application of data sampling approach applied in solving class imbalance in the medical field, more scientist is coming up with combining both under-sampling and oversampling techniques. Thereby, overcoming limitations of a single method with the other.

Cost sensitive learning and ensemble methods are termed as algorithmic-level based approaches i.e., they can modify the classifier's performance with the unbalanced dataset. Cost sensitive methods keep the data used in training unchanged while assigning penalty cost or weights to the misclassification of minority class. The idea of CSL in imbalanced problem is to cause the ML algorithm to pay more attention to minority sample class. Ensemble deals with training a combination of multiple component learners to significantly better the generalization ability of single models.

Ensemble can be seen as a wrapper to other methods [9]. Examples of algorithms incorporated with cost sensitive weighting includes: SVM, ANN, other cost sensitive approaches are SVM_Weight and Weighted ELM (WELM) [9], [52]. Ensemble based models includes EasyEnsemble and Balanced cascade [54].

Asymmetric classifiers, is an addition to the CIL algorithms [54]. They are very similar to the cost sensitive learning algorithms but vary in the way weights are assigned. Unlike cost sensitive, weights assigned to the false negative and false positive samples can be the same. In other words, different weights are not necessarily the aim of the classifier.

3. 4. Data scalability models

The execution of integration analysis on ML models on a single computer has become a lot difficult due to issues like scalability. The performance of data driven models gets better with the constant availability of more and more dataset. However, how can we resolve the issue of big data scalability if performance relies on robustness of the data.

Computationally efficient models proposed for big data scalability includes online ML algorithms, cloud computing and distributed systems for implementation. Here, machine learning and large-scale cloud computing analysis which are the future of multi-omics integration would be briefly discussed.

Online based algorithms such as incremental-decremental support vector machines, Online-sequential extreme learning machine and cost sensitive hinge loss support vector machines are models implemented in addressing scalability (CSHL-SVM) [55]. Online learning algorithms have been so efficient in big data applications because they can train models as the dataset are inputted at once without repeating the process from the beginning, which is literally more optimal, unlike some other learning algorithms which needs the data to be within is reach in memory before training can begin, online algorithms do not need samples to be stored in memory. i.e., it prevents unnecessary space consumption and optimal processing of big data. For instance, OS-ELM (Online-sequential extreme learning machine) is modelled in such a way that it can analyze data has many times as possible without repeating each process when a chunk of samples or a single data is added [9].

Non-iterative algorithms such as random vector functional link (RVFL) [56], extreme learning machines like the dual-layer kernel extreme learning machine (DKELM) [57] . Non iterative models wade off computational complexity as seen in iterative models that requires parameter tuning which can be very exhaustive and time laborious. RVFL is a randomized version of functional link network that can constantly generate weights and keep it fixed during training. Other models still useful in solving scalability issues includes echo state network and liquid state machine [9].

Implementation of distributed systems is another approach in solving data scalability. Here, big data are analyzed on several computers that are connected in a cluster like manner. This enhances computational power to process the data in real-time unlike having a standalone system carry the workload. More so, several parallel machine learning models that are built on MapReduce programming framework and its open-source implementation Hadoop or other variations such as Apache Spark [58], Spark MLlib [59] and Apache mahout [60] have been developed to obtain the desired objective. Examples of algorithms based on MapReduce and spark or singly includes CurboSpark, spark-based parallel SHC algorithm (SHAS), parallel back propagation neural network (PBNN), K-means particle swarm optimization (KMPSO) [61]and a recent addition is the clustering algorithm based on Hadoop known as KAMILA (KAYmeans for MIxed Large data) [62] designed for multiple data types etc.

MapReduce have been employed in a wide variety of supervised, unsupervised, reinforcement learning and deep learning addressing scalability issues because of its fast speed and optimal training time when handling a large number of nodes in multiple data type analysis [63] while Spark is known for its easiness of use and fault-tolerance, thereby improving learning that is required for multiple iterations [64].In addition, other alternatives like vertical scaling approaches such as GPUs, entails the boosting of a single machine's computing power and storage capacity.

Another approach is cloud computing. Why cloud computing to integrate data types or for computationally efficient analysis? Cloud computing is a scalable solution because it offers a wide range of opportunities ranging from the benefits of virtualized resources, containerization, scalable data storage, security, flexible data access and also allows parallelism [65].

Amazon Web Services (AWS) is an open-source cloud computing platform that facilitates the sharing of commonly used datasets stored in the repository. Galaxy cloud allows users to install a private galaxy on AWS and EC2 (Elastic Compute Cloud) [9]. BioVLAB is another freely accessible bioinformatics system deployed on cloud apart from Galaxy. Users have a variety of features that could be manipulated to suite the analysis to be done. BioVLAB-mCpG-EXPRESS, a cloud-based system that accepts three types of raw omics data (gene expression, DNA methylation, and sequence variation) as inputs to do multi-perspective analysis [66]. The system also gives the user multi-level interpretation, allowing the user to interpret the results at each level. Other omics related open-source frameworks that are accessible by many users includes Omics pipe [67], MetaboAnalyst [68] and XCMS online [69].

4. LIMITATIONS OF MACHINE LEARNING-BASED MODEL IN MULTI-OMIC AND CLINICAL DATA INTEGRATION

There has been significant improvement in research based on integrative analysis due to the enormous machine learning models. These significant contributions to the medical field have started making changes in our daily lives. Machine learning applications to many areas include Internet search, speech recognition, product recommendations, image classification, email spam filters [70]. This robust approach has also been implicated in the identification of causal variants associated with a disease trait has shown many successes and researchers are still harnessing the potential of such approaches to multi-omics integrative analysis. The advantages of using machine learning models cannot be overemphasized, as their performance is still gaining attention from biomedical researchers. As earlier stated, several ML algorithms have been employed in tackling multi-omics and clinical data-based challenges. These challenges include high noise, dimensionality problems, overfitting, missingness in the data, scalability issues, analytical variance and are not limited to the aforementioned [71]. This aspect of the study would be focusing on possible limitations related to the robust approaches used in handling biomedical data acquired from diverse modalities.

4. 1. Deep Learning (DL)

DL is an artificial neural network-based machine learning algorithm (ANN). It operates by applying a nonlinear transformation to its input and then using what it learns to generate a statistical model as output [72]. Iterations are repeated until the output meets an acceptable level of accuracy. The label deep was inspired by the number of processing layers through which data must pass. DL is gaining traction as a powerful approach for encoding and learning from heterogeneous and complex data in both supervised and unsupervised settings [72]. In the area of single-omics, multi-omics and biomedical research, deep learning has gained so much traction by researchers as they are well suited to handle complex, heterogeneous and high dimensional dataset such as omics dataset.

4. 2. Problems of Deep Learning in omics integration and biomedical data analysis

Low signal-to-noise ratios, For instance, datasets with unequal proportionality where the samples are small compared to a large number of features or high analytical variance frequently impede omics data analysis [73]. DL algorithms face the challenge of not only analyzing single-omics data but also integrating different types of omics layers [72]. Other sources of information such as medical images data or clinical health records pose a major problem when using deep learning [74], this includes [75]; the volume of the clinical data, the availability of massive amounts of EHR data serves as the foundation for the performance of deep learning neural networks. In the application of DL to the clinical dataset, a good rule of thumb is to have a minimum of 10x of the number of samples as parameters in the network.

More so, In Africa as a case study, patients clinical records are unautomated and access of the masses to hospital facilities would pose a limitation to the amount of available data to train a deep learning model. Furthermore, the quality of the data as input is another challenge, clinical data are less well-structured, unlike omics data. They are highly heterogeneous, ambiguous and incomplete, using such large and diverse datasets to train a deep learning model would be very difficult because of issues like data sparsity, redundancy, and missing values.

The temporality of the data, diseases are constantly evolving unpredictably over time, and existing deep learning models, including those proposed in the medical domain, assume static vector-based inputs, making it difficult to handle the time factor naturally. In addition, a model with so much accuracy and less interpretability **of** the result is regarded as a setback in the clinical domain. Despite their success in a variety of applications, deep learning models are frequently regarded as black boxes. While this may not be a problem in more deterministic areas like image classification, it is critical in the health sector to not only perform quantitative algorithms but also understand why they work. Indeed, the model's interpretability is crucial in determining which phenotypes drive the predictions.

More so, there is increased complexity pertaining to model design and the computing environment required. Although these challenges are still valid because the use of DL methods in omics and precision medicine is a recent field, several scientists are putting in work to resolve the flaws thus improving its purposes. The raising disposal of medical images, clinical health records and as well omics datasets is driving assuring use of deep learning technology, which will play a significant role in this field soon [72].

5. IMPROVED MODELS FOR MULTI-OMIC INVESTIGATION

Some studies have worked in developing new machine learning to investigate several omics data types. New approaches can be built based on two or more algorithms by concatenating them in a manner that can counteract each model's limitations. This part emphasizes a few reformed models designed to integrate multiple omics data.

[76] designed PALM. It works by using a dynamic Bayesian network to reconstruct a unified model and then aligns multiple omics data types. PALM (pipeline for analysis longitudinal multi-omics data) was built to overcome some of the limitations of DBNs such as sampling and progression differences and in the reduction of a cumbersome number of entities and parameters. Although, PALM is a microbiome prediction model, it was implicated in multiple omics types (gene expression, metabolites, microbial taxa) and accuracy in prediction was validated against Baseline model, MTPLasso and MMvec which are microbe-metabolite neural network models.

[77] designed MSCA. It leverages representation-based methods. MSCA (Multiview subspace clustering analysis) was built to overcome the challenges of low-rank representation when applied on multiple data types such as omics because they often assume that the dataset is linear and cannot utilize the geometric features of the original data. MSCA has been implicated using the CCLE dataset of subgroups of tumor cells from multiple origins. MSCA applied local structure in preserving important features in the dataset for effective pattern identification. Here, the performance of the model with SNF, ANF and iCluster+ was evaluated with adjusted rand index score (ARI) at different noise range. The study demonstrated a high rand index score using MSCA than other clustering algorithms, indicating better accuracy with the complex heterogenous datasets due to the combined effect of low rank representation and local structure preservation.

[78] designed ANF. Affinity Network Fusion is a non-probabilistic network that utilizes similarity network fusion but in an upgraded way to counteract some challenges of SNF when using heterogeneous data. For instance, weights can be added to each feature in the multiple datatypes with a clear interpretation while the previous model is always unweighted and gives

a spurious interpretation of the result. ANF applies spectral clustering to reduce noise obtained from non-uniform multiple data types (DNA methylation, miRNA and gene expression). ANF was applied to the TCGA dataset to cluster cancer patients into groups and identification of the cancer subtypes of the patients. The clustering performance of SNF and ANF was evaluated with the adjusted rand index (ARI), p-value of the log rank test and normalized mutual information (NMI). The study demonstrated that ANF performed better than SNF, it still reserved the unstable nature of pair-wise clustering.

[79] proposed the application of Multiview learning which has been applied in other research domains but has not been fully explored in multi-omics data analysis to counteract some challenges such as heterogeneity and noisy nature a common problem in this omics domain. The study designed a framework based on empirical risk minimization know as MV-ERM (Multiview Empirical Risk Minimization). This model was designed to address the overfitting problem when integrating different multitype data. Here, different views or modes represent a fraction of the whole complex processes involved in a biological network. MV-ERM is modeled as an extension of the existing empirical risk minimization principle (ERM) for investigating multiomics data and revealing functional products.

[80] designed a deep learning-based model that is built on the already developed autoencoder known as denoising autoencoder (DAE). The model is an advancement of autoencoders that was constructed to solve dimensionality problem in omics integration. However, with multiple integrations of omics and clinical data types, autoencoders tend to find it difficult to extract informative features because the input and output are equal. To counteract the problem of the existing model, DAE trains the model by imposing noise with the input dataset to extract only features of importance and transposes the data back to its original form by the inbuilt encoder and decoder. To measure the performance of the proposed framework used in learning more robust features, multiple omics ovarian data (mRNA, miRNA and CNV) was analyzed. The study compared the clustering performance of DAE by measuring the silhouette score and log-rank p-values with seven clustering algorithms. DAE had a higher silhouette score and lower p-value showing its significance and performance accuracy.

These models discussed above all performed well based on the different omics data used and as such still needs further validation of performance with other data types for generalizability.

6. CONCLUSIONS

The vast growth of single-cell technologies is promoting the increase in the volume of parameters that can be measured per cell and the number of cells and molecules detected. For this reason, there is more interest in integrating cell data across modalities. The development of multi-omics is encouraging the efforts to build a comprehensive Human Cell Atlas that includes every cell in the human body and that of major model organisms. This will allow subsequent experiments to be performed more quickly and cheaply as information can be transferred from the genome to new data sets through read alignment. An example of a multi-omics method is nanopore sequencing that can sequence RNA and DNA with long reads and can detect nucleotide base modifications. Other multi-omics technologies can detect other biomolecules like proteins and can reveal information about the expression and differentiation of cells. Refinement of multi-omics methods will enable cells to be placed into their spatial context, revealing how cell types differentiate. Thus, different data modalities of single cells within an array of experimental conditions will allow us to move beyond a genome/transcriptome-centric cell view and learn more about the holistic representation of cells. Here, robust approaches such as the diverse machine learning models and even subset of ML, like artificial intelligence has been employed by various researchers in a bid to understand underlying mechanisms associated with a disease trait. Although it is important to know what models to be used before applying it to the collected dataset, because not all models are suitable in investigating multiple omics integration analysis. Regardless of challenges or limitations surrounding this new domain, there are many benefits of multi-omics integrative analysis in implicating genotype-phenotype association in several terminal diseases.

References

[1] S. T. O'Donnell, R. P. Ross, and C. Stanton, The Progress of Multi-Omics Technologies: Determining Function in Lactic Acid Bacteria Using a Systems Level Approach, *Frontiers in Microbiology*, vol. 10. Frontiers Media S.A., p. 3084, Jan. 28, 2020, doi: 10.3389/fmicb.2019.03084

[2] P. S. Reel, S. Reel, E. Pearson, E. Trucco, and E. Jefferson, Using machine learning approaches for multi-omics data analysis: A review, *Biotechnology Advances*, vol. 49. Elsevier Inc., p. 107739, Jul. 01, 2021, doi: 10.1016/j.biotechadv.2021.107739

[3] A. Tebani, C. Afonso, S. Marret, and S. Bekri, Omics-based strategies in precision medicine: Toward a paradigm shift in inborn errors of metabolism investigations, *International Journal of Molecular Sciences*, vol. 17, no. 9. MDPI AG, Sep. 14, 2016, doi: 10.3390/ijms17091555

[4] D. M. Rotroff and A. A. Motsinger-Reif, Embracing Integrative Multiomics Approaches, *Int. J. Genomics*, vol. 2016, 2016, doi: 10.1155/2016/1715985

[5] Z. Y. Yang, Y. Liang, H. Zhang, H. Chai, B. Zhang, and C. Peng, Robust Sparse Logistic Regression with the $(0 < q < 1)$ Regularization for Feature Selection Using Gene Expression Data, *IEEE Access*, vol. 6, pp. 68586–68595, 2018, doi: 10.1109/ACCESS.2018.2880198

[6] C. Wu, F. Zhou, J. Ren, X. Li, Y. Jiang, and S. Ma, A selective review of multi-level omics data integration using variable selection, *High-Throughput*, vol. 8, no. 1, pp. 1–25, 2019, doi: 10.3390/ht8010004

[7] N. J. Mulder *et al.*, Development of Bioinformatics Infrastructure for Genomics Research in H3Africa, *Glob. Heart*, pp. 1–8, 2017, doi: 10.1016/j.gheart.2017.01.005

[8] J. Oyelade *et al.*, Clustering Algorithms: Their Application to Gene Expression Data, *Bioinform. Biol. Insights*, vol. 10, p. 237, Nov. 2016, doi: 10.4137/BBI.S38316

[9] B. Mirza, W. Wang, J. Wang, H. Choi, N. C. Chung, and P. Ping, Machine learning and integrative analysis of biomedical big data, *Genes*, vol. 10, no. 2. MDPI AG, Jan. 01, 2019, doi: 10.3390/genes10020087

[10] B. De Meulder *et al.*, A computational framework for complex disease stratification from multiple large-scale datasets, *BMC Syst. Biol.*, vol. 12, no. 1, p. 60, Dec. 2018, doi: 10.1186/s12918-018-0556-z

[11] N. Rappoport and R. Shamir, Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic Acids Res.*, vol. 46, no. 20, pp. 10546–10562, Nov. 2018, doi: 10.1093/nar/gky889.

[12] Z. M. Hira and D. F. Gillies, A review of feature selection and feature extraction methods applied on microarray data, *Adv. Bioinformatics*, vol. 2015, 2015, doi: 10.1155/2015/198363

[13] L. Wang, Y. Wang, and Q. Chang, Feature selection methods for big data bioinformatics: A survey from the search perspective, *Methods*, vol. 111, pp. 21–31, 2016, doi: 10.1016/j.ymeth.2016.08.014

[14] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, Multi-omics Data Integration, Interpretation, and Its Application, *Bioinform. Biol. Insights*, vol. 14, p. 117793221989905, Jan. 2020, doi: 10.1177/1177932219899051

[15] B. Wang *et al.*, Similarity network fusion for aggregating data types on a genomic scale, *Nat. Methods*, vol. 11, no. 3, pp. 333–337, 2014, doi: 10.1038/nmeth.2810.

[16] M. W. Libbrecht and W. S. Noble, Machine learning applications in genetics and genomics, *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, Jun. 2015, doi: 10.1038/nrg3920

[17] I. Triguero, S. Del Río, V. López, J. Bacardit, J. M. Benítez, and F. Herrera, ROSEFW-RF: The winner algorithm for the ECBDL'14 big data competition: An extremely imbalanced big data bioinformatics problem, *Knowledge-Based Syst.* vol. 87, pp. 69–79, 2015, doi: 10.1016/j.knosys.2015.05.027

[18] J. C. Aledo, F. R. Cantón, and F. J. Veredas, A machine learning approach for predicting methionine oxidation sites, *BMC Bioinformatics*, vol. 18, no. 1, p. 430, Sep. 2017, doi: 10.1186/s12859-017-1848-9

[19] J. Hu, Y. Li, M. Zhang, X. Yang, H. Bin Shen, and D. J. Yu, Predicting Protein-DNA Binding Residues by Weightedly Combining Sequence-Based Features and Boosting Multiple SVMs, *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 14, no. 6, pp. 1389–1398, 2017, doi: 10.1109/TCBB.2016.2616469

[20] Z. Liu, X. Xiao, W. R. Qiu, and K. C. Chou, IDNA-Methyl: Identifying DNA methylation sites via pseudo trinucleotide composition, *Anal. Biochem.* vol. 474, pp. 69–77, Apr. 2015, doi: 10.1016/j.ab.2014.12.009

[21] W. Zhang, T. D. Spector, P. Deloukas, J. T. Bell, and B. E. Engelhardt, Predicting genome-wide DNA methylation using methylation marks, genomic position, and DNA regulatory elements, *Genome Biol.* vol. 16, no. 1, p. 14, Jan. 2015, doi: 10.1186/s13059-015-0581-9

[22] Z.-S. Wei, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, A Cascade Random Forests Algorithm for Predicting Protein-Protein Interaction Sites, *IEEE Trans. Nanobioscience*, vol. 14, no. 7, pp. 746–60, Oct. 2015, doi: 10.1109/TNB.2015.2475359

[23] Z.-S. Wei, K. Han, J.-Y. Yang, H.-B. Shen, and D.-J. Yu, Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests, *Neurocomputing*, vol. 193, pp. 201–212, Jun. 2016, doi: 10.1016/j.neucom.2016.02.022

[24] W. Lin and D. Xu, Imbalanced multi-label learning for identifying antimicrobial peptides and their functional types, *Bioinformatics*, vol. 32, no. 24, pp. 3745–3752, Dec. 2016, doi: 10.1093/bioinformatics/btw560

[25] R. Argelaguet *et al.*, Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets, *Mol. Syst. Biol.*, vol. 14, no. 6, Jun. 2018, doi: 10.15252/msb.20178124

[26] L. De Cecco *et al.*, Integrative miRNA-Gene expression analysis enables refinement of associated biology and prediction of response to cetuximab in head and neck squamous cell cancer, *Genes (Basel).*, vol. 8, no. 1, p. 35, 2017, doi: 10.3390/genes8010035

[27] K. Kira and L. A. Rendell, Feature selection problem: traditional methods and a new algorithm, in *Proceedings Tenth National Conference on Artificial Intelligence*, 1992, pp. 129–134

[28] A. Adorada, R. Permatasari, P. W. Wirawan, A. Wibowo, and A. Sujiwo, Support Vector Machine - Recursive Feature Elimination (SVM - RFE) for Selection of MicroRNA Expression Features of Breast Cancer, in *2018 2nd International Conference on Informatics and Computational Sciences (ICICoS)*, Oct. 2018, pp. 1–4, doi: 10.1109/ICICOS.2018.8621708

[29] M. B. Kursa, A. Jankowski, and W. R. Rudnicki, Boruta – A System for Feature Selection, *Fundam. Informaticae*, vol. 101, no. 4, pp. 271–285, 2010, doi: 10.3233/FI-2010-288

[30] F. Santosa and W. W. Symes, Linear Inversion of Band-Limited Reflection Seismograms, *SIAM J. Sci. Stat. Comput.*, vol. 7, no. 4, pp. 1307–1330, Oct. 1986, doi: 10.1137/0907087

[31] H. Zou and T. Hastie, 'Regularization and variable selection via the elastic net', *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 67, no. 2, pp. 301–320, 2005, doi: 10.1111/j.1467-9868.2005.00503.x

[32] C. Meng, O. A. Zeleznik, G. G. Thallinger, B. Kuster, A. M. Gholami, and A. C. Culhane, 'Dimension reduction techniques for the integrative analysis of multi-omics data', *Brief. Bioinform.*, vol. 17, no. 4, pp. 628–641, Jul. 2016, doi: 10.1093/bib/bbv108

[33] Q. Mo, R. Shen, C. Guo, M. Vannucci, K. S. Chan, and S. G. Hilsenbeck, 'A fully Bayesian latent variable model for integrative clustering analysis of multi-type omics data', *Biostatistics*, vol. 19, no. 1, pp. 71–86, 2018, doi: 10.1093/biostatistics/kxx017

[34] J. C. Costello *et al.*, 'A community effort to assess and improve drug sensitivity prediction algorithms', *Nat. Biotechnol.*, vol. 32, no. 12, pp. 1202–1212, Dec. 2014, doi: 10.1038/nbt.2877

[35] G. Tini, L. Marchetti, C. Priami, and M. P. Scott-Boyer, 'Multi-omics integration-A comparison of unsupervised clustering methodologies', *Brief. Bioinform.*, vol. 20, no. 4, pp. 1269–1279, 2018, doi: 10.1093/bib/bbx167

[36] C. Dimitrakopoulos *et al.*, 'Network-based integration of multi-omics data for prioritizing cancer genes', *Bioinformatics*, vol. 34, no. 14, pp. 2441–2448, Jul. 2018, doi: 10.1093/bioinformatics/bty148

[37] C. J. Vaske *et al.*, 'Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM', *Bioinformatics*, vol. 26, no. 12, Jun. 2010, doi: 10.1093/bioinformatics/btq182

[38] A. Hosseini, T. Chen, W. Wu, Y. Sun, and M. Sarrafzadeh, 'Heteromed: Heterogeneous information network for medical diagnosis', *Int. Conf. Inf. Knowl. Manag. Proc.*, pp. 763–772, 2018, doi: 10.1145/3269206.3271805

[39] Y. Zhang, A. Li, C. Peng, and M. Wang, 'Improve Glioblastoma Multiforme Prognosis Prediction by Using Feature Selection and Multiple Kernel Learning', *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 13, no. 5, pp. 825–835, Sep. 2016, doi: 10.1109/TCBB.2016.2551745

[40] A. Rakotomamonjy, 'SimpleMKL', vol. 9, pp. 2491–2521, 2008.

[41] J. Barretina *et al.*, 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012, doi: 10.1038/nature11003

[42] Y. Zhang *et al.*, 'Integrative functional genomics identifies regulatory genetic variant modulating RAB31 expression and altering susceptibility to breast cancer.', *Mol. Carcinog.*, vol. 57, no. 12, pp. 1845–1854, Dec. 2018, doi: 10.1002/mc.22902

[43] N. Aben, D. J. Vis, M. Michaut, and L. F. A. Wessels, 'TANDEM: A two-stage approach to maximize interpretability of drug response models based on multiple molecular data types', in *Bioinformatics*, Sep. 2016, vol. 32, no. 17, pp. i413–i420, doi: 10.1093/bioinformatics/btw449

[44] R. Shen, A. B. Olshen, and M. Ladanyi, 'Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis', *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, Nov. 2009, doi: 10.1093/bioinformatics/btp543

[45] K. A. Hoadley *et al.*, 'Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer', *Cell*, vol. 173, no. 2, pp. 291-304.e6, Apr. 2018, doi: 10.1016/j.cell.2018.03.022

[46] Q. Mo *et al.*, 'Pattern discovery and cancer gene identification in integrated cancer genomic data', *Proc. Natl. Acad. Sci. U. S. A.*, vol. 110, no. 11, pp. 4245–4250, Mar. 2013, doi: 10.1073/pnas.1208949110

[47] J. Barretina *et al.*, 'The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity', *Nature*, vol. 483, no. 7391, pp. 603–607, Mar. 2012, doi: 10.1038/nature11003

[48] T. D. Nguyen, T. Tran, D. Phung, and S. Venkatesh, 'Latent patient profile modelling and applications with mixed-variate restricted Boltzmann machine', in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and*

Lecture Notes in Bioinformatics), 2013, vol. 7818 LNAI, no. PART 1, pp. 123–135, doi: 10.1007/978-3-642-37453-1_11

[49] M. Liang, Z. Li, T. Chen, and J. Zeng, 'Integrative Data Analysis of Multi-Platform Cancer Data with a Multimodal Deep Learning Approach', *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, vol. 12, no. 4, pp. 928–937, 2015, doi: 10.1109/TCBB.2014.2377729

[50] M. Kim, I. Oh, and J. Ahn, 'An Improved Method for Prediction of Cancer Prognosis by Network Learning', *Genes (Basel).*, vol. 9, no. 10, p. 478, Oct. 2018, doi: 10.3390/genes9100478

[51] L. Zhang *et al.*, 'Deep Learning-Based Multi-Omics Data Integration Reveals Two Prognostic Subtypes in High-Risk Neuroblastoma', *Front. Genet.*, vol. 9, no. OCT, p. 477, Oct. 2018, doi: 10.3389/fgene.2018.00477

[52] L. Nanni, C. Fantozzi, and N. Lazzarini, 'Coupling different methods for overcoming the class imbalance problem', *Neurocomputing*, vol. 158, pp. 48–61, Jun. 2015, doi: 10.1016/j.neucom.2015.01.068

[53] V. H. Barella, E. P. Costa, and A. C. P. L. F. Carvalho, 'ClusterOSS : a new undersampling method for imbalanced learning', *Brazilian Conf. Intell. Syst.*, pp. 1–6, 2014

[54] L. Nanni, C. Fantozzi, and N. Lazzarini, 'Coupling different methods for overcoming the class imbalance problem', *Neurocomputing*, vol. 158, pp. 48–61, 2015, doi: 10.1016/j.neucom.2015.01.068

[55] B. Gu, X. Quan, Y. Gu, V. S. Sheng, and G. Zheng, 'Chunk incremental learning for cost-sensitive hinge loss support vector machine', *Pattern Recognit.*, vol. 83, pp. 196–208, 2018, doi: 10.1016/j.patcog.2018.05.023

[56] L. Zhang and P. N. Suganthan, 'A comprehensive evaluation of random vector functional link networks', *Inf. Sci. (Ny).*, vol. 367–368, pp. 1094–1105, 2016, doi: 10.1016/j.ins.2015.09.025

[57] T. V Nguyen and B. Mirza, 'Dual-layer kernel extreme learning machine for action recognition', *Neurocomputing*, vol. 260, pp. 123–130, 2017, doi: 10.1016/j.neucom.2017.04.007

[58] M. Zaharia *et al.*, 'This open source computing framework unifies streaming, batch, and interactive big data workloads to unlock new applications', *Commun. ACM*, vol. 59, no. 11, 2016, doi: 10.1145/2934664

[59] X. Meng *et al.*, 'MLlib: Machine Learning in Apache Spark', *J. Mach. Learn. Res.*, vol. 17, pp. 1–7, 2016

[60] R. Anil *et al.*, 'Apache Mahout: Machine Learning on Distributed Dataflow Systems', *J. Mach. Learn. Res.*, vol. 21, pp. 1–6, 2020

[61] M. Sherar and F. Zulkernine, 'Particle swarm optimization for large-scale clustering on apache spark', *2017 IEEE Symp. Ser. Comput. Intell. SSCI 2017 - Proc.*, vol. 2018-Janua, pp. 1–8, 2018, doi: 10.1109/SSCI.2017.8285208

[62] A. H. Foss and M. Markatou, ‘kamila: Clustering mixed-type data in R and hadoop’, *J. Stat. Softw.*, vol. 83, pp. 1–44, 2018, doi: 10.18637/jss.v083.i13

[63] A. L’Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz, ‘Machine Learning with Big Data: Challenges and Approaches’, *IEEE Access*, vol. 5, no. April, pp. 7776–7797, 2017, doi: 10.1109/ACCESS.2017.2696365

[64] P. Gupta, A. Sharma, and R. Jindal, ‘Scalable machine-learning algorithms for big data analytics: a comprehensive review’, *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 6, no. 6, pp. 194–214, 2016, doi: 10.1002/widm.1194

[65] I. A. T. Hashem, I. Yaqoob, N. B. Anuar, S. Mokhtar, A. Gani, and S. Ullah Khan, ‘The rise of “big data” on cloud computing: Review and open research issues’, *Inf. Syst.*, vol. 47, pp. 98–115, 2015, doi: 10.1016/j.is.2014.07.006

[66] M. Oh, S. Park, S. Kim, and H. Chae, ‘Machine learning-based analysis of multi-omics data on the cloud for investigating gene regulations’, *Briefings in Bioinformatics*, vol. 22, no. 1. Oxford University Press, pp. 66–76, Jan. 01, 2021, doi: 10.1093/bib/bbaa032

[67] K. M. Fisch *et al.*, ‘Omics Pipe: a community-based framework for reproducible multi-omics data analysis’, *Bioinformatics*, vol. 31, no. 11, p. 1724, Jun. 2015, doi: 10.1093/BIOINFORMATICS/BTV061

[68] J. Chong *et al.*, ‘MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis’, *Nucleic Acids Res.*, vol. 46, no. Web Server issue, p. W486, Jul. 2018, doi: 10.1093/NAR/GKY310

[69] E. M. Forsberg *et al.*, ‘Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online’, *Nat. Protoc.*, vol. 13, no. 4, pp. 633–651, Apr. 2018, doi: 10.1038/nprot.2017.151

[70] E. G. Dada, J. S. Bassi, H. Chiroma, S. M. Abdulhamid, A. O. Adetunmbi, and O. E. Ajibuwa, ‘Machine learning for email spam filtering: review, approaches and open research problems’, *Heliyon*, vol. 5, no. 6, p. e01802, 2019, doi: 10.1016/j.heliyon.2019.e01802

[71] J. Fan, F. Han, and H. Liu, ‘Challenges of Big Data Analysis.’, *Natl. Sci. Rev.*, vol. 1, no. 2, pp. 293–314, Jun. 2014, doi: 10.1093/nsr/nwt032

[72] J. Martorell-Marugán *et al.*, ‘Deep Learning in Omics Data Analysis and Precision Medicine’, in *Computational Biology*, Codon Publications, 2019, pp. 37–53

[73] D. Grapov, J. Fahrmann, K. Wanichthanarak, and S. Khoomrung, ‘Rise of Deep Learning for Genomic, Proteomic, and Metabolomic Data Integration in Precision Medicine’, doi: 10.1089/omi.2018.0097

[74] D. Ravi *et al.*, ‘Deep Learning for Health Informatics’, *IEEE J. Biomed. Heal. Informatics*, vol. 21, no. 1, pp. 4–21, Jan. 2017, doi: 10.1109/JBHI.2016.2636665

[75] R. Miotto, F. Wang, S. Wang, X. Jiang, and J. T. Dudley, ‘Deep learning for healthcare: Review, opportunities and challenges’, *Brief. Bioinform.*, vol. 19, no. 6, pp. 1236–1246, May 2017, doi: 10.1093/bib/bbx044

[76] D. Ruiz-Perez *et al.*, 'Dynamic Bayesian Networks for Integrating Multi-omics Time Series Microbiome Data', *mSystems*, vol. 6, no. 2, Apr. 2021, doi: 10.1128/msystems.01105-20

[77] Q. Shi, B. Hu, T. Zeng, and C. Zhang, 'Multi-view subspace clustering analysis for aggregating multiple heterogeneous omics data', *Front. Genet.*, vol. 10, no. JUL, p. 744, Aug. 2019, doi: 10.3389/fgene.2019.00744

[78] T. Ma and A. Zhang, 'Integrate multi-omic data using affinity network fusion (ANF) for cancer patient clustering', in *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Nov. 2017, vol. 2017-Janua, pp. 398–403, doi: 10.1109/BIBM.2017.8217682

[79] N. D. Nguyen and D. Wang, 'Multiview learning for understanding functional multiomics', *PLOS Comput. Biol.*, vol. 16, no. 4, p. e1007677, Apr. 2020, doi: 10.1371/journal.pcbi.1007677

[80] L. Y. Guo, A. H. Wu, Y. X. Wang, L. P. Zhang, H. Chai, and X. F. Liang, 'Deep learning-based ovarian cancer subtypes identification using multi-omics data', *BioData Min.*, vol. 13, no. 1, pp. 1–12, Aug. 2020, doi: 10.1186/s13040-020-00222-x