Topological Data Analysis and Multi-Scale Biomedical Signal Processing: AI-Enhanced Systems for Complex Biomedical Data Analysis and Pattern Extraction

Dr. Joseph Msabila, Associate Professor of Information Systems, University of Nairobi, Kenya

1. Introduction to AI in Biomedical Data Analysis

Recent developments in AI have played a significant role in revolutionizing different areas of biomedical data analysis. Modern high-throughput assays allow the generation of the most comprehensive description of the living cell at the individual molecular level on a grand scale. Large-scale omics data generation techniques have not only allowed the development of detailed and accelerating knowledge about genes, gene products, metabolites, and interaction networks. In addition, molecular phenotyping has entered the experimental domains of advanced biology. However, with the acquisition of such groups of data, it has also become apparent that we need new, advanced, and quite different experimental and analytical tools that can help assemble the complex information contained in the new datasets, spanning genomics, transcriptomics, proteomics, and metabolomics.

Moreover, the field of computational biology, bioinformatics, and systems biology has greatly evolved during the past decade, and powerful tools have emerged that could help us use laboratory omics data to a fuller extent in order to make testable hypotheses. These methods promise a deeper biological understanding of living systems by linking molecular functions, system structure, and context. We focus on a hybrid combination of experimental and computational tools that could be employed to obtain mechanistic insights about complex biological systems in the age of the cellular network. Revolutionary developments in in silico modeling, such as Boolean network analysis, differential equation-based modeling, and agent-based models, are now available for a plethora of in vitro studies, and many modeling and data integration training programs have also been launched as modeling and simulation curricula.

In this paper, a compilation of selected emerging evolutions towards an understanding of the organizing principles of the cell in humans as a networked system with immediate relevance for personalized medicine is presented. This paper is segmented into five sections, with further illustrative details on the advanced learning, highly studied networks, and subjective or relative research preferences as developed during the past few years. The first part is focused on integrative biology, bioinformatics, and systems biology as a type of "network biology" or "integrated medicine" or "integrated systems medicine" or "integrated physiology" and network medicine, where networks could be symbolic, interactomic, gene expression-based, disease-related, or metabolic-related. In this section, we have assembled 18 subsections, with various topics; some of this content reflects the research abstracts and focuses of millions of researchers and scientific scholars who might have explored disease comorbidities or gene co-expression and regulation of the interactome, tissue-specific, systems-related functionalities therapeutically and behaviorally through knowledge and information on signals and systems in the networks.

2. Fundamentals of Genomics, Proteomics, and Metabolomics Data

In genomics, the aim is to study the complete set of DNA sequences, which is called the genome of an organism. Therefore, large-scale analysis methods have been developed to determine the order of DNA bases in the entire genome or for the coding part of the genome. In this latter case, they provide the genetic information extracted from mRNA. Therefore, they are called transcriptomics or often referred to as genomics. Analyzing DNA provides fundamental information in the understanding of many complex or multifactorial traits, or so-called hereditary diseases.

In proteomics, the intention is to understand the function of proteins; in other words, the structure of the proteins in the proteome of an organism. Proteins play several roles in the cell and reflect the activity of a certain mechanism. In particular, they might point out the disease mechanism. In a broad sense, proteomics aims to acquire knowledge about the entire profile of proteins, which is called the proteome. Proteomics often refers to the study of proteins as such, in particular for their structure and function. Proteomics might possibly measure what is happening in the cell, whereas genomics measures the potential. Metabolomics refers to the study of the complete set of small molecule metabolites found in a cell. They reflect, to some degree, the impact of genotype and

karyotype, in close interaction with the environment. These profiles are varying and show the activity and state of an organism at the gene and protein level.

These levels and fields (genomics, proteomics, and metabolomics) are tightly connected to each other. In particular, metabolites are products of DNA, RNA, and protein activity. In contrast, protein activity results in changes of gene expression after genome interpretation. Hence, they represent a downstream effect of gene and protein activity and reflect the determined phenotype. Metabolites directly reflect the outcome of cell activity and bear important information about phenotype, regardless of the genetic or proteomic background.

3. Machine Learning Techniques for Biomedical Data Analysis

Machine learning techniques play a crucial role in understanding and analyzing complex biomedical data. One of the simplest machine learning techniques is supervised learning, in which models learn the input-output relationship from the labeled dataset. Domain knowledge is essential in choosing features and oversampling techniques. Unsupervised learning algorithms are used to detect patterns, group similar data points, and make sense of situations where labels are not provided. While minute changes in a response are difficult to capture, several supervised learning classification and regression algorithms can be employed. Random forest, support vector machine, artificial neural network, and logistic regression are algorithms of choice to study biological populations for effective prediction results. Feature engineering is the most crucial step in selecting vital features for oversampling and undersampling techniques.

Other machine learning techniques like deep learning, generative adversarial networks, and hybrid models are used widely to automate and support medical experts in rational decision-making. Deep learning can capture complex data structures and perform high-level data abstraction. Deep learning applications include disease detection, severity measurement, and treatment decision processes. Convolutional neural networks are widely used deep learning models for effective feature extraction. A combination of techniques for better performance results in disease diagnosis has been proposed. The integration model of the support vector machine, principal component analysis, and particle swarm optimization establishes a method for predicting datasets on various diseases. Long short-term memory combined with has been shown to effectively predict

diseases using medical images. These methods show an improved mean score as compared to individual methods.

3.1. Supervised Learning Algorithms

Supervised learning. Any detection of structure, extraction of meaningful biomarkers, or automatic decision-making using an AI-enhanced system needs labeled data (i.e., one for which either the identity of the classes is known or the expected outcome is characterized). Supervised learning is a process of associating input features with an output by learning from a dataset, rather than through explicit implementation of the input-output mapping. Algorithms used for supervised learning are aimed at finding models that can optimally associate input features with the appropriate output label. Multiple types of supervised learning methods are available, such as linear regression, decision trees, random forests, support vector machines, neural networks, and so on. Each method has its own advantages and disadvantages; algorithms favored for some specific applications may not be so for others. Therefore, selecting an algorithm for a specific task in the biomedical domain is crucial. In general, supervised learning methods can be used for classifying diseases, predicting outcomes from a set of transcriptomic and/or genomic features, detecting certain attributes, and developing computer-aided diagnosis systems, among others.

Numerous performance evaluation metrics are available for quantifying the predictive accuracy of AI-enhanced models. Normalized accuracy along with other metrics such as area under the receiving operating characteristic curve, precision, recall, and an F1-score are frequently used. Additionally, classification models can be validated using multi-fold cross-validation methods, which protect against many known pitfalls, such as model overfitting and biased performance estimation. In a study using gene expression data, the performance of ten different machine learning algorithms in various applications, such as predictive modeling of disease classification, outcome prediction, and outcome classification, was assessed. Although an optimal predictive model was not readily available among the algorithms tested, six different models that achieved a stable performance in the external validation dataset were identified. It was concluded that gene expression-based model development for clinical outcomes is possible using diverse algorithms and feature selection techniques as long as the sample size is sufficiently large.

3.2. Unsupervised Learning Algorithms

Unsupervised learning algorithms mainly aim to find the hidden patterns within biological datasets by extracting useful knowledge. Unsupervised learning deals with the problem of knowing neither input nor output, where special emphasis is given to exploiting the dataset's internal structure. In this subsection, we will mainly focus on clustering techniques and dimensionality reduction techniques used in unsupervised classification. Here, we do not have labeled outcomes that guide learning, so in clustering and dimensionality reduction, these are complex scenarios. These scenarios are used to find discoveries based on the data itself. These automated molecular subgrouping results help us find the subpopulations. The critical components are genetics and the non-negative background expression values of mRNA normalized from the data. The K-means algorithm is improved by soft versions, as well as using non-Euclidean distance and correlation-based measures. It includes self-organizing maps, singular value decomposition maps, and other cluster analysis methods such as agglomerative hierarchical clustering and conventional dendrograms.

There are many applications of these unsupervised methods such as PCA and hierarchical clustering in the domain of genetics or genomics, by combining gene expression datasets to find which genes are responsible for the observed phenotype. Other than this, these methods can be used to find and rectify proteomic plasma metabolite differences between subjects with phenylketonuria versus a control group. Clustering might be used in various steps of data analysis as a discovery tool or as a hypothesis-generating method. For instance, in the application domain of genomics and proteomics, one of the major goals is to identify a robust and unbiased way to group relevant subpopulations. Associations between these groups and, for instance, patient survival or response to treatment can then be evaluated. However, clustering generates a large number of features, and choosing a single sample once the clustering process is completed may be arbitrary. This brings us to the question of how to choose an optimal number of clusters and how to interpret the results. The selection of a large number of clusters will separate the dataset from its original composition, while selecting a very small number of clusters may bring too little information. The outcome of the clustering can be used to randomize the clusters and to visualize the relations between the samples and the features.

3.3. Deep Learning Models

While machine learning performs 'learning' using a large number of training datasets, deep learning, specifically, is a part of a subset of machine learning that performs classification, regression, and clustering. As opposed to shallow learning, which uses only a single hidden layer for complex computations, deep learning uses multi-layered artificial neural networks to perform computations. A major advantage of deep learning is its ability to automatically 'extract' or 'learn' instrumental features of interest from raw data without any manual intervention. These deep learning models have been demonstrated to be efficacious in raw time series multi-dimensional data for complex pattern recognition. A few important deep learning architectures include the convolutional neural network, which has proven to be successful in analyzing or depicting an image on a pixel-by-pixel basis, and the recurrent neural network architecture and its variants, which are effective in the processing of high-dimensional data sets such as sequential or time series data.

Deep learning architecture is best suited if the model must perform decision-making based on learning patterns from high-dimensional data that may contain redundant and irrelevant information. Real-life practical scenarios usually exhibit high-dimensional data properties in the field of biomedicine or other fields within. Biological data such as genomics, transcriptomics, and proteomics profiling data in the post-genomic era are common high-dimensional datasets with many features and a small number of samples. However, some open research issues and challenges should be addressed in the case of biomedical big data. Personalized analysis is difficult if the available individual data are limited. Application to integrated data analysis requires input data with multiple new feature vectors, samples, time, space, and scale. Computing, memory, and storage resources require expensive hardware. Currently, the biomedical field shows a paradigm shift due to the potential applications for deep learning on NGS and medical imaging. In genomics, many studies have been introduced, such as expression modeling, gene regulation, multi-omics integration, and computational methods. With the revolution of deep learning methods in medical imaging, most of the top dealers have released algorithms for automatic detection, classification, risk assessment, and prediction-related diseases that use CNN. In practice, the successful implementation of deep learning methods is still in the early stages of research. The number of available cases is still limited, and reproducibility and robustness are still challenges.

4. Challenges and Limitations in Analyzing Biomedical Data

AI approaches for analyzing biological data have gained traction in recent years. However, these approaches do have several limitations, often remaining "hype" in more complex biomedical applications. One of the key issues relates to the quality of medical data to which we want to apply AI methodologies. Due to the very soft "protection" barriers between systems in living organisms, data recorded from them is affected by noise, missing values, and inconsistencies to a larger degree compared to typical engineering datasets. Moreover, data collection in medicine requires more ethical considerations, and the data-driven models may inherit biases from the data to which they have been trained. There are often non-trivial algorithmic challenges in analyzing these datasets, and there are computational infrastructure costs in using many AI models, which is often difficult to afford. In this section, we elaborate more on the challenges of analyzing biomedical/clinical data with respect to the AI methodologies mentioned in the previous sections. We start with the data challenges: noise, missing values, data quality, high dimension, etc. The next subsection discusses the ethical aspects of data collection, the compliance issues faced while using these datasets for analytics, and the governance issues. Subsequently, we deliberate the need for the expensive computational infrastructure requirements when dealing with the AI techniques we mentioned, estimating quality models and large-scale omics data. The other challenge is interpretability. Then we explain the regulatory aspects in biomedical data, the approval times for clinical use, and the compliance requirements needed to test and use the AI models with biomedical data. Finally, we share real-world use case examples that suffer from these challenges to the advent of AI.

5. Case Studies and Applications in Genomics, Proteomics, and Metabolomics

Artificial intelligence (AI) techniques have shown promise in enabling synergistic interdisciplinary analysis using diverse data types—including genomics, proteomics, and metabolomics—to address real-world biomedical research problems. Here, we use a series of case studies to illustrate work at the interface of advanced AI methodologies and professional practice in diverse areas, ranging from clinical diagnostics to the determination of functional characteristics in biological molecules and processes. We demonstrate how this integration of AI with traditional biological analyses streamlines the way in which people work and enables them to address questions that would not have been feasible to tackle using a conventional unimodal approach. Case studies in

genomics highlight their potential use in personalized medicine endeavors: AI-enriched gene selection algorithms can be used to suggest biologically meaningful markers and aid in patient stratification in cancer. Integration of protein-protein interaction data with significant variants provides new disease-stratifying candidate biomarkers. AI embedding can also significantly boost applications in proteomics using sequence-based tools for functional annotation of proteins. Computational mass spectrometry-based metabolic profiles can also be harnessed using AI in combination with proteomics.

The variability of biomedical and genomic data can be exploited to define new subpopulations of patients and stratify the disease using computational biology tools enhanced by AI models. AI powering can help to define new strategies for treatment and personalized medicine, representing the future of pharmaceutical research. In this review, we present some of the pioneering applications of AI in proteomics and genomics as well as some of the most informative and promising computational methods in the field of metabolomics. We further introduce the challenges encountered in this scenario from the perspective of biologists, geneticists, and bioinformaticians. Furthermore, intrinsic difficulties in the use of black-box AI models for robust biomedical profiling are addressed and lessons learned are provided.

6. Future Direction

With the rapid development of AI algorithms and computational capabilities, various biomedical problems will be solved by AI-aided systems in the future. We are starting to see some emerging trends. With improvements in AI algorithms, nearly all incurable diseases might have heterogeneous molecular features encoded in multi-omics data. Progress regarding both traditional solutions and innovative methodologies will be adopted for biomedical interest. Predictive, prognostic, and therapeutic development for patients suffering from those diseases will thus be the focus of medical study. Furthermore, with personalized medicine evolving as a manageable new part of medical practice, we will see more developments in this area. This extension will tackle big data based on individual preferences and healthcare promotion rather than pathology assessment. With the progressive construction of nationwide and worldwide biobanks and genome databases for subsequent generations, prevention rather than treatment in medical care will be highlighted.

Additionally, relevant issues of AI-enhancement therapies, such as iatrogenic diseases for those who receive unhealthy reports based on AI algorithms rather than personal intercommunications and physical examinations, privacy-violated problems, continuous and risk-free technical upgrades, and on-time on-demand systems will be explored for critical improvements. Early disease prevention and superb personalized medical service construction will be the focus of scientific research. AI-aided research will develop into the fields of integration of multi-omics data, functional drug discovery, individualized strategies, extension of personalized medicine lists, current and previous disease assessment, the interaction of medical, hereditary, and autoimmune factors, and so on. The development of clinical assessment from disease to health to a complete picture will eventually redefine itself. Thus, several research directions are suggested, and this will enlighten and inspire some interdisciplinary researchers and practitioners to step forward for further developments in this area. In addition, we highlight that interdisciplinary collaboration is crucial in this area since current methodologies are inflexible and computationally time-consuming. Many improvements in this area are impressive for application, especially in individualized precision medicine. Ethical considerations should include those for the medical field, requiring focus on disease prevention and health care. Thus, medical responsibility is highlighted as the frontier point to guide ethical considerations in this area. Additionally, transparent cooperation in ethical AI-enhanced disease prevention and health management is illustrated in four key issues.

7. Conclusion

In this work, we thoroughly discussed various AI-based methodologies and tools used in the analysis of complex biological and biomedical information. We demonstrated the extendibility of these methodologies to other bioinformatics research fields, such as genomics, proteomics, or systems biology. AI and machine learning have revolutionized the way biomedical data are analyzed, generating new knowledge through the integration and prediction of complex data types. We proposed that this revolution is just the beginning and characterized it as the first stage to interpret and conquer the multidimensional and inter-related information extracted from biological and medical environments. We also argued for the necessity of advanced analytical methodologies to address these challenges. However, although all these advancements are promising, several unresolved issues still exist in the AI- and ML-driven analysis of biomedical and

biological data. This includes the adaptation of models to new scenarios, the reduction of possible sources of bias, integration of different data types to perform trans-omic analysis, and refining models to provide better performances on larger datasets. Investigating these issues has been proposed as a strategy to drive future research and help make AI an essential component of new research scenarios.

Another important aspect concerns the application of AI models to the practical context to demonstrate that these tools may assist healthcare professionals in their daily work. This requires close collaboration between computer scientists and end-users, professionals that are expert in the healthcare and clinical use of specific methodologies. Moreover, interaction with the clinical community is an essential component of the methodology design and needs to be included from the data exploration stage to secure the necessary support to develop and apply the AI tool. We presented an application example in breast cancer using gene expression data. In the example, the AI model was intended to suggest an adjuvant therapy for a subgroup of patients recruited within a clinical trial. The definition of supplementary therapy was made through straightforward and scalable AI application in a real clinical research scenario. We showed that the trained model outperforms the consensus classification performance for the identification of very-high-risk patients and suggested that a new class consisting of patients might be considered: high-responding patients with chemotherapy. This result may open a window to personalize the treatment of patients with chemotherapy and guide the treatment of very-high-risk patients, in whom drugs are ineffective.

Although AI and ML are crucial for the future of healthcare, there are other aspects which need to be addressed, such as further investigating ethical issues regarding the use of personal information and issues related to privacy, vulnerability of the community, and data protection. How can the potential of AI be harnessed to change human care? This can be achieved by continuous education and research through conferences, courses, and forums to make healthcare stakeholders aware of the power of AI and show how new diagnostic modalities are helped by AI. Open dialogues between healthcare and industry organizations, governments, and societies are also crucial for the widespread adoption of AI technology for health-related applications on an international scale. Overall, the future of AI applied to healthcare looks promising and

indicates the need for continuous research to maximize the potential of this technology, making a positive impact on healthcare systems worldwide.