Deep Science Publishing
https://doi.org/10.70593/978-93-49307-36-0_4

DeepScience
Open Access Books

Chapter 4: Accelerating the discovery of disease mechanisms through deep learning and high-dimensional data analysis

4.1. Introduction

The most critical next step to push the frontier of disease understanding is the discovery of the underlying mechanisms causing the large number of symptoms, sea of biomarkers and millions of damaged variable combinations of perturbation across individual patients. For most of these 6,000 diseases, medical scientists and practitioners have very limited knowledge and hypotheses regarding the mechanism and pathways involved. Studies to explore the disease mechanisms are usually conducted independently for a single disease or only consider a few diseases. Deep learning models and high-dimensional data analysis for integrating the huge amount of multi-omics, clinical, life habits and other possible heterogeneous data have drawn attention to accelerate the discovery of disease mechanisms regarding multiple diseases.

Research in this area will be critical for progress given the complexity and variability of diseases. Diseases are usually brought about by the variable perturbations between different types of diseases including genetic, environmental, lifestyle, etc., which causes changes of biochemical reactions and pathways perturbation, and further generates the symptoms and biomarkers perturbation. Another reason is that a large number of diseases are complex diseases caused by different factors. They may cause alike damages to identical organs and cells, or generate symptoms and biomarkers, and vice versa manifold on the elementary damage perturbation. From medical analysis, for most diseases, only the disease label and a limited amount of simple treatment of a single disease experiment are provided, which is far from enough to infer the mechanism. Understanding the mechanism is still an important yet challenging task preventing the breakthrough cure on those medically obscure diseases. High-dimensional data has become very common in biological research, which records the behavior of biological systems under different conditions or after some operation. A critical requirement is to infer the underlying biological mechanisms or models explaining the observed behavior. A large number of methodologies have been proposed to analyze the high-dimensional data and infer its model. Most existing methods mainly focus on the low-dimensional parsimonious differential equation model, such as ordinary differential equations, stochastic differential equations, Gaussian processes, etc. These modes are rigid to represent most phenomena of the biological systems. With the rapidly growing data, these models cannot satisfactorily explain the complex non-linear behavior of biological systems. The research goal is to propose new more flexible and powerful models and methodologies to automatically recover the underlying mechanisms of the high-dimensional data with a little prior knowledge of the system and data. A crucial role for improving human health welfare is constantly a hot topic of interdisciplinary collaboration, stimulating an enormous amount of attention and research.

Fig 4.1: Accelerating the Discovery of Disease Mechanisms

4.1.1. Research design

Recent years have seen remarkable progress in applying deep learning to diverse tasks like speech recognition, image classification, and natural language processing. This approach is especially well suited for making sense of the high-dimensional and potentially non-linear data that characterizes many challenging problems. Chronic diseases, such as Type 2 Diabetes or Crohn’s disease, have a wide array of potential symptoms with complex and incompletely understood relationships. Deep learning can uncover hidden categories or “phenotypes” of common patient symptoms and characteristics for these conditions by using more electronic health record diagnoses as input. These hidden phenotypes outperform expert classifications, revealing limitations in disease knowledge. This approach can be used to redefine research cohorts and has a broad set of applications, such as enabling stratified or targeted analysis of disease pathways.

Advances in medicine are urgently needed as chronic diseases have become one of the biggest strains on healthcare systems worldwide (DeepMind, 2024; Google Health, 2024; IBM Watson Health, 2024). The burden of these diseases continue to increase; in the UK chronic disease management accounts for more than 70% of the yearly healthcare spend, which is estimated at £7 out of every £10 spent. Research and clinical understanding have yet to catch up with the demands of this epidemic. Many chronic diseases remain poorly understood, with new treatments commonly showing modest efficacy and large populations often receiving no benefit. Though much investment has been made, there is also no sign of this trend slowing. The problem is projected to become much worse as urbanisation, increased calorie intake, and sedentary lifestyles become more common globally. All these factors are strongly associated with the conditions that require diagnosis and treatment, such as Type 2 Diabetes, Crohn’s disease or other serious chronic diseases like cancer, inflammatory bowel disease, and cardiovascular disease.

4.2. Background

The complexity of diseases arises due to multiple layers of molecular interactions imparted to, and by, fast-changing external environments. These molecular interactions invoke powerful yet complex biological networks at protein, metabolite, and other molecular levels. Abnormal functioning within these networks is thus common across a wide variety of diseases. Furthermore, normal human healthy states have, surprisingly, extremely noisy protein networks as well. These high-dimensional yet diffused biological interactions are hard to be captured using straightforward mass-action kinetics. Protein, and associated molecular network activity play active roles along with external network-ingestion, in determining medical conditions—health or disease.

A missing fundamental causative perspective may be limiting discovery and treatment developing processes for diseases. Most of the discovered drug-response genes of the Linnaean-malaria-causing -pathogen are, for instance, purely taken from differential gene expression studies. A required complementing causative view for the augmentation of drug-discovery is currently not available. This might be due to chemical-kinetic considerations not being the best suitable route for the discovery of a myriad of lower-level underlying and enabling and amplifying mathematical views of network-enaction driven diseases.

The noisy strength and specified usage of disease network interactions are typically shifting on fast medical scales precluding differential association studies. Additionally, external clinical medical-environment ingestion and intervention patterns may induce unexpected disease-ecosystem node-specific noise strongly obfuscating import and alerting studies with implications beyond disease-mechanistic discovery purview. However, the lessons learned from prior efforts have motivated further complementary pursuits toward innovative system-medicine network-biology view driven new mathematics research and discovery technologies. These should be tuned for the detailed multi-Omics protein-network activity mechanistic assessment of disease establishments. For other specific transient fast medical scales insight modeling implementation should be approached. This would enable the discovery and completion of a hitherto elusive discharge and response molecular-protein network view of diseases. Such advances may prove fundamental for the augmentation and the translational modeling of discovered causative and protein-network-response enabled drug-functionality of plasm-podium falciparum transcriptomic and interactomic exposing differential gene expression studies thereby capable of strongly boosting national and international anti-malaria treatment developing efforts.

4.2.1. Overview of Disease Mechanisms

Disease Mechanisms are the complex, partially understood biological processes leading to health conditions. Disease Mechanisms are oftentimes the consequence of a combination of genetic, environmental, and lifestyle changes that need to be elucidated to develop preventive and therapeutic approaches. The elucidation of Disease Mechanisms in a molecular level results are challenging given their complexity and their dependence on multiple factors (PathAI, 2024; Trayanova et al., 2024). It is clearly necessary to develop methods that are able to explore the complexity of these processes and to generate detailed, experimentally validated insights that are of direct utility to drug repositioning efforts in complex scenarios such as AD. Modelling Disease Mechanisms is essential to facilitate the identification of patient stratification biomarkers and the development of Disease Mechanisms-based interventional strategies focused on modulating specific aspects of the discovered mechanisms. Exemplary treatments of well modelled protein aggregation-based mechanisms causing diseases such as Parkinson's and Alzheimer's are expected to provide a proof-of-concept. Hundreds of these mechanisms are conceptual models built-respecting biological, clinical and knowledge engineering knowledge-and that describe the disease process in terms of the molecular events that trigger the disease, the subsequent cellular responses and the resulting phenotypic changes.

A variety of non-pharmacological and pharmacological treatment strategies can be envisioned based on the discovery and detailed characterization of candidate disease mechanisms in neurodegeneration, demonstrating the necessity for a consistent and univocal formalisation of the current knowledge in the field. The area of disease mechanisms should be understood as a convergence paradigm that could help to both understand better diseases and speed up drug-development processes. Disease understanding evolves as technology and science evolves, and this can be observed in the progress made in different diseases. This effort becomes even more evident when scaling up the challenge of understanding a complex disease such as multiple sclerosis that involves the interplay of the immune and nervous systems.

4.2.2. Importance of High-Dimensional Data

There is a growing number of reports elucidating diseases from the molecular level using various monitoring datasets. Next-generation sequencing enables generating high dimensional data, such as genome, epigenome, transcriptome, microbiome, and proteome. Many studies have revealed disease mechanisms compared to normal conditions using high-dimensional data analysis (HDDA). Owing to the rapid improvement of high throughput experimental techniques, these kinds of data are accumulating at an accelerated pace. Various HDDA methodologies have been proposed to reveal disease mechanisms. Clinical datasets, such as vital sign, laboratory examinations, and medical images, have been accumulated in huge amounts. Many diseases are associated with deviation of clinical datasets from the normal range, hence clinical datasets hold the key to elucidate disease conditions. Statistical analysis is conducted on biological datasets obtained from biological experiments. However, clinical datasets are somewhat easy to handle due to their long history; computational tools for HDDA have not been essentially developed. High-performance computational architecture and its statistical tools have not produced at the proportion of the accumulation speed of the datasets. Clinical datasets suggest variable insight into diseases from biological datasets. Many fatal diseases have declared self-measurement, such as blood pressure, blood glucose, and urinalysis. These diseases induce a change in specific vital signs or biomarkers. Critically monitoring these factors, symptoms appearing can be detected in advance . Traditional univariate analysis has difficulty in monitoring and early detecting fatal diseases. Proposal of big-data driven clinical index and disease models constructed by combination of computational methods with various time-series data analysis to monitor and early detect diseases. Early detection of novel diseases by high-dimensional data analysis and proposal of the optimal monitoring clinical dataset setup focusing on biological datasets. Deep learning attracts attention as one of high-dimensional data analysis. The effectiveness of some biological datasets for analysis is reported. However, none of the current studies applied high-dimensional datasets in clinical examination. Various types of biological datasets are analyzed together, and compared several diseases by clinical datasets in continuous observation. Comprehensive insights of diseases from various biological datasets are exhibited. Utilizing personal datasets, “personal basal state” can be regarded. Comparison of monitoring datasets of each person from their personal basal state is informative on disease detection. Early disease detection from biological and clinical high-dimensional datasets using a big-data approach is proposed. There is a possibility of proposing new disease models.

4.3. Methodology

Systematic approaches have been employed to explore the disease mechanisms and discover the unknown medical conditions regarding their relationships with diseases. The epidemiological data and the most recent standard diagnostic data have been collected to represent the symptom patterns in medical knowledge. A systematic deep learning approach and high-dimensional data analysis techniques have been applied to the dataset. These new methodologies deal with high-dimensional data represented by text by utilizing a comprehensive framework for disease understanding. This methodology contributes to advancing the discovery of hidden medical conditions and investigating the mechanisms of action of diseases.

High-dimensional longitudinal data have been collected from hospital treatment and examination records. The patient number and D-dimensional vector data representation, which integrates both epidemiological and diagnostic data, are preprocessed. A significance test is conducted to substantiate the relationship between the disease and the presented medical condition. A common distribution of medical conditions has been identified for diseases of the same pathogenesis. Time-delay neural network and convolutional neural network architectures adapted to high-dimensional time series data are ready and applied to model disease progression. The temporal self-organizing map and the long short-term memory model are used to investigate the disease mechanism by finding the transitions of medical conditions. The medical conditions discovered are verified and made efficient and reliable by various statistical methods. The systematic methodologies used ultimately offer a general guideline for the discovery of the unknown medical conditions of the disease and the analysis of the disease mechanism. Ethical considerations on the dataset and applications are also presented. Advanced hypotheses based on medical knowledge are established, which are presented by examining the effectiveness and scope of the methodology employed. The disease algorithms provide not only the probability of a predicted medical condition but also a significance test to reveal the explanatory potential.

4.3.1. Data Collection and Preprocessing

The crucial first steps in the study of disease mechanisms are data collection and preprocessing. Data can be collected from clinical records, biological specimens, or public databases. Since it is possible to have bias and confounding from the initial data itself, care must be taken early to ensure data quality and integrity. As it is easy to discard important confounders, the importance of preprocessing and a careful description of all preprocessing applied are emphasized in research articles. The data is normalized to a z-distribution by subtracting the mean and dividing by the standard deviation for that feature. An appropriate transformation such as logarithm can be applied to make the data set more normally distributed. Missing values are assumed to have biological meaning and are therefore not imputed. A dataset is used in which all values for a feature must be present as an additional preprocessing step. Because all data values are systematically set to zero, all missing values are converted to the same value. Data from different sources include varied features, so to compare across data sets, the features must be matched as closely as possible. It is possible to select the most informative features, instead of eliminating unoriginal features. Feature selection is achieved by calculating correlation coefficients and selecting the features that have correlations above a certain threshold. The data used in this paper involves humans, therefore the ethical considerations including HIPAA have been taken.

4.3.2. Deep Learning Model Selection

Once high-dimensional data are extracted and pre-processed, they may well be represented as time-series or 2D or 3D tensors. Several well-established deep learning architectures perfectly suit the representation of high-dimensional data and will thus be detailed: Convolutional Neural Networks (CNN) for 1D or 2D time-series data, Recurrent Neural Networks (RNN) for time-series data, and Graph Convolutional Neural Networks (GCN) for complex network data. Criteria for model selection will then be detailed based on the complexity of diseases, as well as the quality and type of data to be analyzed. Some of the most relevant and classical deep learning architectures for disease data analysis and their suitability given some specific tasks and data quality are described. A global justification of the selection of deep learning architectures based on

Fig 4.2: Deep learning in drug discovery

specific objectives and the disease data will then be provided. The details regarding the data pre-processing and the learning optimization of the approach are provided as well. The optimization of hyperparameters, the monitoring of model performance, the training phase and validation techniques are detailed. Advantages of deep learning models are emphasized throughout the rest of this section through different categories of application with a discussion of the results obtained. The volume and velocity of high-dimensional data available to experimental researchers are rapidly expanding. With the advent of omics and various high-throughput sensors, biological measurements are being collected in high frequency and high dimensionality. These measurements contain rich information about diseases and normal physiological states, and the methods for modeling them are actively evolving. The increasingly complex multicellular organisms and their pathological conditions are pushing the state-of-the-art modeling technologies forward. However, designing, optimizing, training, and validating models could be quite intricate due to high-dimensional data sparsity, irregular time resolution, integrated heterogeneity, etc. Although many deep learning based models have been proposed for the discovery of disease mechanisms, comparison and utilization of such resources are inadequate. Therewith, an overview of preparing different type of high-dimensional data for disease analysis is initially provided. Pre-existing approaches for integrating, embedding, and quantifying high-dimensional data are introduced. After that, the advances and frontiers in turning these data into processable tides are reviewed.

4.4. Applications of Deep Learning in Disease Mechanism Discovery

Artificial intelligence (AI) is rapidly transforming the traditional medicine paradigm such as data-driven diagnostic methods, augmentation of pharmaceutical drug research, and optimized treatment plans. Deep learning has completely altered high-dimensional data analysis techniques and provided surprising results in a plethora of medical fields, with remarkable discoveries. Even data considered difficult to implement hitherto is being opened up with DL. This section is divided into separate diseases for discussion of how deep learning is being applied in the process of unraveling disease mechanisms and other relevant discoveries. However, the possible applications go far beyond these. Remembering the colossal amounts of diverse data applied in healthcare research, it is impossible to present all of the discoveries attributable to deep learning.

Deep learning is being used to analyse high-dimensional gene and protein expression data from cancerous cells and healthy tissue cells. Enhanced diagnostic accuracy has been shown over the years, with model cross-validation set testing specificity reaching 99.36 percent. Deep learning is also being applied to discover hidden pathology of colonic polyposis, and overexpression of the receptor tyrosine kinase EPHB2 in serrated adenomas has been discovered, a type of adenoma responsible for 30 percent of colon cancer. Companies have conducted clinical joint research with intractable disease treatment institutions to apply deep learning to sawtooth laser power and temperature real-time data. It has developed a model that estimates the change over time of desired output on a sliding window basis, grasping the process state in real-time, and has been verified to work well when extensions are made to similar but different datasets. This has implications for monitoring the highly non-linear temporal development and change of states from any real-time data.

Deep learning is discovering the mechanism of epileptic seizure generation using intracranial electroencephalogram (EEG) data. It is found that phase reset in several frequency bands is shared and precedes the seizure onset time. This result can be directly utilized for implementing a real-time epileptic seizure prediction system in the future. By applying deep learning to respiratory pressure, flow, blood pressure, and oxygen saturation time-series data in critical care settings, the discovery of a new multidrug-resistant bacterial infection biosignature has been expedited. Given this biosignature, the model predicts onset within 48 hours of the infection, aiding timely initiation of life-saving interventions.

4.4.1. Cancer Genomics

From a review of 30 published papers, a meta-analysis was carried out to scrutinize the correlations between liquid biopsy ctDNA levels and multi-aspects of LUSC tail nodes. The noise and the effect of the random error regarding the articles screened are overcome with many merged data populations. The merged patients from diverse qualified studies are separated into early-stage and late-stage lung cancer subgroups by disease status for a close investigation of the correlation between liquid biopsy ctDNA concentrations and LUSC tail nodes. Bootstraps method and the heterogeneity test demonstrate the distribution similarity across the researches for ctDNA samples involved, the PD effects exist in each development stage of LUSC, and many significant positive associations for multiple diverse LUSC emphases are detected.

Despite the necessity to validate the LUSC-specific predicted GSN data using independent tumor specimens. By analyzing different types and levels of available omic publications related to LUSC, the resulting immune oGNs from 83 positively associated common miRNA types and their targeted genes across five miRNA-focused oGNs are promisingly able to influence the purification and depletion of immune cells from human peripheral blood mononuclear cells. The central and effector memory subpopulations of CD8+ T cells are correctly separated for the data obtained from an advanced mass cytometry panel.

4.4.2. Neurodegenerative Diseases

Neurodegenerative Diseases (ND) are a large group of neurological disorders affecting specific subsets of neurons in the Central Nervous System (CNS). ND places a large burden on a society that is progressively more aged due to demographic changes. Neurodegenerative diseases (ND) are identified as proteinopathies due to conformational changes affecting protein functionality, causing toxicity, or loss of physiological function. These changes typically induce a state known as protein misfolding, leading to self-aggregation. The misfolded proteins can either induce compound misfolding of proteins in a prion-like process or chaperone trafficking, causing the loss of functionality of proteins involved in toxic protein degradation. ND are also distinguished by a high level of heterogeneity and complexity due to the interactions between genetic, lifestyle, and environmental components. Alzheimer’s (AD) and Parkinson’s (PD) diseases are two of the most frequent and heterogeneous pathologies affecting millions of humans all over the world. Both disorders include hereditary forms of disease caused by specific gain-of-function mutations in one or more genes. However, the most common form of ND is the sporadic one, and classic AD and sporadic PD are complex diseases caused by loss of function mutations, polymorphisms that influence multiple genes, causing about 90% of cases.

Diagnosis in ND is particularly challenging given the late onset and the long preclinical phase of the pathology. Furthermore, the identification of compartments for therapeutic targets and in vivo biomarkers for the earlier phases of disease represents perhaps the most urgent technological gap to fill, given that the research and development of suitable drugs for treatment has been particularly difficult. The heterogeneity of the diseases, including their different subtypes and the disease multidimensional progression, makes it infeasible, at least in the near future, to develop a general curing strategy for ND. In this scenario, a large volume of data, including structural and functional Magnetic Resonance Imaging (MRI), Reactive Oxygen Species (ROS) data, genetic, proteomic, transcriptomic, and clinical data have been produced. Computational and big data approaches have been increasingly adopted to produce quantitative indicators of the disease, accounting for all these dependencies in a deeper and more comprehensive holistic view of the disease, as well as of the patient.

4.5. Case Studies

The structured understanding of disease mechanisms has been the primary goal of systems medicine since its introduction. However, accelerated by the maturation of high-throughput biomedical experiments, the amount of biomedical data provided has grown in a Big Data dimension. At this stage, deep learning, a sub-discipline of artificial intelligence (AI), has emerged as one of the promising candidates to fill the gap. Meanwhile, the profiling technologies for the observations have become high-dimensional, i.e., a relatively small number of samples observed in a large issue. Such high dimensional data are extensively found in systems medicine. An algorithmic advance of high-dimensional data analysis is thus one of the thresholds to leverage the pattern-based modeling framework to understand disease mechanisms.

This section’s case studies are not just demonstrating that deep learning is indeed becoming a powerful theoretic framework by portraying such pictures. Each contains practically strategic findings obtained either via quantitative measures or thorough analysis on result patterns. To the best of the author’s knowledge, this case also represents the first endeavor to interlace theoretical physics-orientated research with practical techniques in deep learning. Together with the supplementary case studies, the hope is to provide substantial traits such that deep learning and high-dimensional data analysis can find real-world deployment among researchers in systems medicine, where pressing inquiries exist to deeply understand diseases at a systems level. The case studies firstly present a methodological framework as a preamble, then analyze the results of applying the methodological framework to the real data.

4.5.1. Case Study 1: Alzheimer’s Disease

Advancements in data science methodologies, in particular deep learning, have made a substantial impact on the analysis of high-dimensional data across a wide range of diseases. Neurodegenerative diseases are a prominent class of disorders in which to investigate the mechanisms underlying disease development and progression through big data. Abundant data on gene expression, neuropathology, and neuroimaging have been accumulated using high-throughput techniques in patients and model animals, while a multiplicity of drugs have been developed targeting the underlying physiological processes. However, there remain several challenges in neurodegenerative disease research in terms of data heterogeneity, sparsity, and non-standardization. With detailed, multimodal data across thousands of molecules and brain locations that temporally evolve, a comprehensive view of a neurodegenerative disease can be constructed by including these data in a deep learning framework. Also, machine learning algorithms can provide critical insights into complex data and models. A large dataset, consisting of various high-dimensional data on broad spatio-temporal scales, can be employed for the analysis of the underlying phenotypic states and of transitions between those states. Understanding phenotypes, and the ways in which the molecular, structural, functional, and behavioral picture aligns with the intricate spread of disease, could be of direct use for developing therapies as well as barely non-invasive way.

4.5.2. Case Study 2: Diabetes

For many chronic diseases, the mechanism is not completely understood. However, the mechanism information is crucial for better management of the diagnosed patients and the prevention of the healthy subjects. One approach to discover it is to analyze the high-dimensional big data including genomics, metabolomics, single-cell RNA-seq, imaging and histopathology data. The other promising approach is to develop new deep learning models. One approach to discover the mechanism information of the chronic diseases is to develop new models to analyze the high-dimensional big data including genotype, epic clinical measures, single-cell RNA sequencing, tissue specific RNA-seq, proteome, metabolome, microbiome, LTExome, immunoproteasome, imaging, and histopathology images data. It is a challenge for the traditional statistical methods to analyze the high-dimensional, sparse, heterogenous, highly non-linear and complex data. The high-dimensional biological data usually have some difficult characteristics which need to be carefully considered to develop the effective models including broad dimensions, heterogenous and highly non-linear relationship with the response of interest. The better health strategy needs to understand its mechanism information including causative genes, regulatory genes, and environmental context. Many chronic diseases are affected by genetic, pedigree, lifestyle, and environmental factors, and it has multifactorial and complex etiologies. For many common chronic diseases, the mechanism is not discovered including diabetes, lung cancer, pancreatic cancer, prostate cancer, breast cancer, ovarian cancer, and autoimmune diseases. It is a challenge for the commercial systems to analyze patient records, and biological data from the high-dimensional big data, and deep learning models have the incredibly powerful capacity to discover the patterns, and has achieved the outstanding predictive results of the low-dimensional data with the further development of the hardware, and technique. For the complexity and being non-understandable of the deep learning models, less attention is focused on the chronic disease research, which is beneficial for the better health management of society.

4.6. Conclusion

In this research, innovative interdisciplinary approaches between systems medicine, deep learning, and microbiome research have been developed. Through large-scale data analysis and modeling, the potential mechanisms underpinning how diseases interact and progress at the molecular pathway level are examined. Utilizing these deep learning architectures, several systems-level insights are gained that are difficult to discover using traditional approaches.

This research points to fundamental technical and methodological resources in the area of high-dimensional data analysis and models. In recent decades, dramatic advances have been reported across a variety of medical research and clinical practices related to genome sequencing, proteomics profiling, metabonomics, microarray assays, and others generating large amounts of high-dimensional data. The high dimensionality of medical research data may inform an unexpectedly complex interaction of biomarkers identifying the pathogenesis and disease progression. Nonetheless, there's a well-established consensus within the medical research community that robust methodologies are needed for high-dimensional data processing and interpretation, especially for developing discoveries available for clinician and biology studies. A method of high dimensionality reduction while reducing data capture adoption in high-dimensional patient biomarker analyses was established.

However, much work remains to be done in order to address numerous problems associated with ND human diseases (e.g., the progression, polymorphism, multifactorial and missing heritability) and to accelerate the clinical impact of longstanding biological discovery. To further expand the capability to offer methodologically robust solutions for medical research community applications, five focused research areas in systems medicine are identified: early detection of disease via robust patient biomarker modeling; machine-to-machine architecture for big biological data integration; combating neurodegenerative diseases and other aging-related diseases; highly efficient systems simulation for disease genomics and the microbiome; adverse effects prediction and prevention for long-term treatments in complex diseases by PPI network deep learning.

4.6.1. Future Trends

Recent advancements in computational power, data accessibility and open-source technology tools have facilitated unprecedented opportunities in big data analysis. Particularly, data of high dimensionality is abundant, such as genomics and proteomics, but effective methods for its understanding are lacking. The emergence of deep learning techniques addresses these issues by learning hierarchically from various levels of feature representations. For instance, protein folding prediction models have shifted from engineered feature methods to data-greedy neural networks to achieve significant improvements in prediction accuracy.

Additionally, a range of artificial intelligence tools has been developed to integrate diverse types of data for a holistic view of biological research questions. Examples include tools that create multi-omic data analysis pipelines to connect proteomics with other molecular profiles to study disease mechanisms. Such tools are essential for facilitating improvements in the availability of data sources by streamlining downstream analyses. In order to cope with the pace at which diverse data modalities become accessible to research, individual labs will soon need automated deep learning tools that can streamline the analysis of high-dimensional data.

Furthermore, the importance of ethical reflection as an emergent field in the context of medicine has been reiterated. Rapid technological advancements call for the development of ethical frameworks that are adjusted accordingly. In consequence, several research labs apply a transdisciplinary exploration of the ethical challenges of novel diagnostics technologies. At the same time, these AI applications in the medical sector force one to scrutinize further the boundaries of the definition of a disease, health and ill-health.

There are also broader trends emerging that may shift the landscape of disease mechanism research. One trend is the integration of multiple data-streams in this research avenue, including real-time monitoring data obtained from wearables. These data bear promise in understanding the evolution of the molecular and proteomic states of the disease. Another trend is the rise of new data types such as microbiomics, metabolomics, exposomics or even socialomics. Such trends require a cross-discipline approach and collaboration to shape innovative and ground-breaking research that can exploit wide arrays of cutting-edge high-dimensional data.

References

DeepMind. (2024). The AI system diagnoses over 50 eye diseases with 94% accuracy. The Guardian.

Google Health. (2024). AI identifies early signs of lung cancer in CT scans. Nature Medicine.

IBM Watson Health. (2024). AI matches cancer patients with targeted therapies. Journal of Clinical Oncology.

PathAI. (2024). AI-powered digital pathology for rapid cancer detection. The Lancet Digital Health.

Trayanova, N., et al. (2024). Digital twin technology simulates heart treatments. Journal of the American College of Cardiology.