Explore Medical Science and Global Health；Vol.1 2025
E-ISSN: 3079-1790 P-ISSN: 3079-1782
DOI: https://doi.org/10.70267/emsgh.2501.1015
Published by: Zeus Press

Application of Deep Learning for Early Disease Diagnosis and Biomarker Discovery

Praise Ye*

College of Natural & Agricultural Sciences, University of California, Riverside, Riverside, 92507, United States of America

**Corresponding author：Praise Ye, E-mail: praise.yep@gmail.com.*

Abstract

Major diseases such as cancer, neurodegenerative diseases and cardiovascular diseases have had serious impacts on the global public health system. Early diagnosis is key for improving treatment outcomes, reducing mortality rates and alleviating the socioeconomic burden. With the rapid development of technologies such as high-throughput sequencing, single-cell omics and spatial transcriptomics, biomedical research has entered a new data-driven stage. How to effectively mine the key information related to the early stage of disease from these complex and high-dimensional multiomics data has become the core issue of current research. This article systematically reviews the research progress of deep learning in the early diagnosis of diseases and the discovery of biomarkers. First, the basic principles of deep learning and its advantages in processing biomedical data were introduced. Subsequently, its typical applications in transcriptomics, proteomics, single-cell and spatial omics, as well as multiomics integrated analysis, were expounded. Meanwhile, the potential value of deep learning in noninvasive detection, such as liquid biopsy, was discussed. The results show that deep learning can automatically extract key features from complex biological data and identify early disease signals that are difficult to detect via traditional methods, providing a new technical approach for disease prediction and precise diagnosis. However, issues such as data heterogeneity, insufficient interpretability of the model, and obstacles to clinical translation remain the main factors restricting its wide application. Future research should focus on enhancing model transparency, sharing high-quality data, and establishing interdisciplinary collaboration mechanisms to accelerate the clinical application and promotion of deep learning in the field of precision medicine.

Keywords

deep learning, biomarkers, early diagnosis, multiomics analysis

1. Introduction

Major diseases, such as cancer, neurodegenerative diseases, and cardiovascular diseases, pose severe challenges to the global public health system and impose heavy social and economic burdens. Among various coping strategies, early diagnosis is regarded as the key to improving treatment outcomes, reducing mortality rates and alleviating the pressure on medical resources. Biomarkers play a core role in this process, enabling the prediction, classification and dynamic monitoring of diseases. However, traditional biomarker discovery methods often have limitations such as high noise, a single and single-dimensional perspective, and difficulty in capturing dynamic and complex changes during the occurrence and development of diseases, which limits their sensitivity and specificity in early diagnosis.

Moreover, with the rapid development of technologies such as high-throughput sequencing, single-cell omics, and spatial transcriptomics, a vast amount of multimodal biomedical data has emerged. These data offer unprecedented opportunities for a deeper understanding of disease mechanisms at the molecular level, but they also pose significant challenges to traditional bioinformatics analysis methods. Against this backdrop, artificial intelligence, especially its important branch, deep learning, has demonstrated tremendous application potential. Deep learning is based on a multilayer neural network structure and can automatically learn abstract features and inherent patterns from high-dimensional and complex data, avoiding the limitations of traditional methods that rely on manually designed features. It is particularly suitable for mining potential weak signals related to the early stage of disease from massive amounts of omics data.

This review aims to systematically identify and explore the research frontiers of how deep learning promotes the early diagnosis of diseases and the discovery of biomarkers. This article first briefly introduces the basic concepts of deep learning and its unique advantages in this field. Then, it elaborates on in detail the specific applications and representative achievements of deep learning in transcriptomics, proteomics, single-cell and spatial omics, and multiomics integration. In addition, its prospects in noninvasive tests such as liquid biopsy are reviewed. In addition, an objective analysis is conducted on the current challenges in terms of data, models, and clinical transformation. Finally, future research directions and clinical transformation paths are needed to emphasize the importance of interdisciplinary cooperation in promoting the transition of this field from theoretical research to clinical practice.

2. The Foundation of Deep Learning in Biomarker Discovery and Early Diagnosis

2.1 What is deep learning? Why is it suitable for this field?

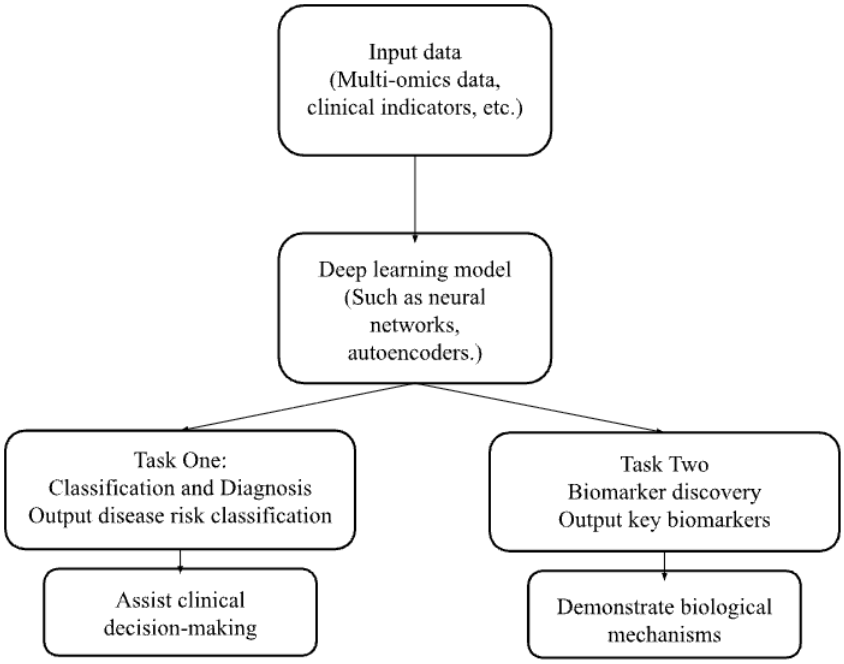
Artificial intelligence is a science that endows machines with intelligent behaviors, and deep learning is a key technology for achieving artificial intelligence. Deep learning imitates the neural network of the human brain and uses multilayer neural networks to learn the patterns in data. Unlike traditional methods that require experts to manually select features from data, the main advantage of deep learning lies in its ability to automatically learn and extract useful features from raw data. For example, when dealing with a large amount of gene expression data, there is no need to prespecify which genes are important. Deep learning models can independently discover which gene combination patterns can distinguish healthy individuals from early-stage patients. This automatic learning ability makes it particularly suitable for processing complex and high-dimensional data generated by modern biomedicine, thereby helping to discover weak early disease signals that are difficult to identify via traditional methods.

2.2 What can deep learning mainly do in early diagnosis?

In the process of early disease diagnosis and biomarker discovery, deep learning mainly accomplishes two core tasks, as shown in Figure 1.

Classification and diagnosis. This is the most direct application. Deep learning models can predict new samples after massive amounts of data are learned, similar to intelligent classifiers. For example, after blood test data are input, the model can directly output the risk probability of early-stage cancer, providing an auxiliary diagnostic basis for clinical practice. In addition to being used for diagnosis, the model can also be used to discover new biomarkers. Some deep learning models, such as autoencoders, are adept at reducing dimensionality and extracting features from complex data, identifying the key features that best represent the essence of the data. By analyzing these key features that the model considers, related molecules, such as specific genes or proteins, can be inferred in reverse, thereby identifying potential novel biomarkers at an early stage of the disease.

Figure 1. Schematic diagram of the task of deep learning in early diagnosis and biomarker discovery.



3. Application of Deep Learning to Omics Data

With the popularization of high-throughput sequencing technology, omics data have become the foundation of disease research. Deep learning, with its ability to handle high-dimensional data, shows great potential in mining such data for early diagnosis.

In traditional transcriptomics and proteomics, deep learning is widely applied to identify complex patterns of gene expression and protein–protein interactions. For example, the DeepSEA model can predict how variations in noncoding DNA regions affect the binding of transcription factors, thereby helping to understand the role of these seemingly useless sequences in the early stages of disease. The DeepBind model can predict the interactions between proteins and DNA/RNA, revealing potential disease regulatory pathways. Unlike traditional differential representation analysis, deep learning models such as autoencoders and transformers can directly learn and compress the most critical features from massive overall data in an unsupervised or weakly supervised manner. This approach is more likely to identify weak signals and novel biomarkers that vary among different patients but generally indicate the early state of the disease, thereby avoiding the biases that traditional methods may cause.

Single-cell technology and spatial transcriptomics can reveal the heterogeneity and spatial location information of cell populations, and deep learning models are key tools for analyzing such complex data. For example, scVI and scGPT can effectively integrate multiple single-cell datasets, precisely identify different cell subtypes, and infer the trajectories of cell development or disease evolution, that is, the process by which cells transition from one state to another. This is crucial for understanding the subtle changes in specific cell populations (such as immune cells) in the early stages of a disease. By analyzing samples from the disease prodromal stage, these models can compare the transcriptomes of specific cell types with differences in health status, thereby identifying the earliest molecular abnormalities and providing targets for ultraearly diagnosis and intervention.

A single type of omics data can reflect only a certain aspect of life activities, whereas the integration of multiple types of data can more comprehensively depict disease maps. Deep learning provides an ideal platform for achieving such integration. Models such as DeepMOFA can integrate data from different sources, such as transcriptomics, methylomes, and proteomes, into a unified framework to learn feature vectors that represent the overall biological state of samples. This integration can reveal disease characteristic patterns that cannot be detected at the individual omics level. Furthermore, deep learning can combine multiomics features

with patients' clinical indicators (such as age and imaging results) to construct more powerful predictive models. For example, this type of fusion model can more accurately classify early subtypes of cancer, thereby providing a basis for personalized treatment.

4. Application of Deep Learning in Liquid Biopsy and Noninvasive Detection

In addition to directly analyzing tissue or cell samples, deep learning also has great potential in noninvasive detection fields such as liquid biopsy, which is highly important for achieving convenient large-scale early disease screening. Liquid biopsy acquires disease information by analyzing trace substances such as circulating tumor DNA, exosomes or circulating RNA in the blood. In the early stage of disease, the content of these signals in the blood is extremely low, and traditional methods have difficulty achieving accurate detection. Deep learning models, especially convolutional neural networks (CNNS, which are good at recognizing image styles and can be used for one-dimensional sequence or signal pattern recognition) and recurrent neural networks (RNNS, which are good at processing sequence data), perform well in identifying weak biological signals from complex noises. This type of model can effectively distinguish weak mutations or abnormal molecular patterns related to early-stage cancer or Alzheimer's disease from background noise, thereby significantly improving the detection sensitivity.

In addition, the application of deep learning has also expanded to metabolomics and respiratory gas analysis. The substances produced by human metabolism or the components in exhaled breath can reflect the health condition of the body. When the disease first occurs, these components undergo subtle changes. Deep learning can analyze these complex data, capture the characteristic changes related to early lesions, and provide noninvasive early warning methods for diseases such as cancer.

The greatest advantage of combining deep learning with these noninvasive detection technologies lies in their clinical feasibility. Operations such as blood drawing or collecting exhaled breath are simple, safe and cause little pain, making them easy to repeat and suitable for large-scale promotion and census among the general population. By conducting convenient screening for a large number of people, individuals at high risk of disease can be identified more effectively, achieving early detection and intervention, which is highly valuable for improving public health.

5. Challenges and limitations

Although deep learning has shown great potential in the biomedical field, it still faces challenges in three aspects: data, models, and clinical transformation in practical clinical applications.

First, data are one of the biggest bottlenecks. Biomedical data typically exhibit significant heterogeneity. Data from different hospitals, testing platforms, or experimental batches show systematic differences, which may lead to a model performing well on one dataset but experiencing a significant decline in accuracy on the other datasets. More crucially, the high-quality data used for training the model, especially the early disease samples precisely labeled by experts, are not only limited in quantity but also costly to obtain, which directly restricts the training effect and generalization ability of the model. Furthermore, patient data involve highly sensitive private information, and their use and sharing are strictly restricted. This makes it extremely difficult to collect multicenter and large-scale datasets, and big data are the prerequisite for deep learning to leverage its advantages.

Second, the model itself has inherent limitations. When the number of early disease samples is limited, complex models are prone to bias and may only remember some irrelevant details or noise in the training data, without truly understanding the diagnostic information provided by the data. This situation reduces the credibility and adoptability of the model in clinical scenarios. Furthermore, models developed for specific groups of people (such as residents of a certain area) may experience a decline in performance when used in other groups, indicating that the adaptability and stability of the models still need to be improved.

Finally, the transformation of research results into clinical tools faces considerable challenges. Models that perform well in the laboratory need to be repeatedly tested and adjusted in a complex clinical environment and effectively integrated with the existing hospital information system. This process is complex and time-consuming. In addition, medical products are subject to strict regulatory oversight, involving procedures such

as medical device approval, algorithmic fairness, and ethical review. Any of these procedures may pose an obstacle to implementation. If the research methods are not unified and the data and code are not made public, other scientists will be unable to replicate and verify the experimental results, which will hinder the healthy development of the entire field.

6. Future Outlook

Despite numerous challenges, the application prospects of deep learning in the field of early diagnosis remain broad. Deep learning offers new tools and methods for early disease diagnosis. Its core advantage lies in its ability to automatically identify early disease signals that are difficult to detect with traditional methods from massive and complex biological data. However, it should be recognized that there is still a long development path for this technology to move from the laboratory to clinical application. Its progress not only depends on the innovation of the model structure but also requires coordinated advancements in multiple aspects, such as data quality, model interpretability and practical application.

The key to future research lies in solving several practical problems: first, enhancing the interpretability of the model to enable clinicians to understand the diagnostic basis, thereby increasing trust and adoption of the model; second, addressing the difficulties in obtaining high-quality medical data and the limitations in data sharing; and third, enhancing the adaptability of the algorithm across different groups of people and devices to prevent performance degradation of the model in practical applications.

Successful medical artificial intelligence projects rely on interdisciplinary collaboration. Biologists need to understand the principles of algorithms, computer scientists need to master clinical needs, and doctors should be involved in the entire process of model development and validation. This deep integration of multiple disciplines is the core driving force for the development of the field and the realization of clinical application.

The application of deep learning in the medical field offers a new perspective: through data and algorithms, it is expected to identify disease signals earlier, achieve timely intervention and prevention, and thereby enhance public health levels. In future research, both technological innovation and insights into real-world problems should be taken into account to ensure that research outcomes have practical significance and clinical value.

7. Conclusion

In conclusion, deep learning technology, with its powerful feature extraction and pattern recognition capabilities, provides unprecedented tools for the early diagnosis of diseases and the discovery of biomarkers. This technology can mine potential key features related to early lesions from high-throughput omics data, medical images and noninvasive detection data, significantly enhancing the ability to recognize and intervene in the early stages of disease.

However, to transform the technological potential of deep learning into practical clinical value, the core goal lies in achieving deep integration across disciplines. Computational science, biology and clinical medicine need to break down domain barriers and form a synergy. Computational scientists need to have a deep understanding of real clinical scenarios, while doctors and biologists should also actively participate in the process of model construction and validation.

In the future, efforts should be made to promote the standardization of data collection and processing, facilitate the opening and sharing of high-quality data while safeguarding patient privacy, and continuously carry out validation research oriented toward clinical transformation. Only through these joint efforts can the revolutionary role of deep learning in the era of precision medicine be fully leveraged to achieve the goal of effectively reducing the burden of major diseases and improving human health.

References

Mandair, D., Reis-Filho, J. S. and Ashworth, A. (2023). Biological insights and novel biomarker discovery through deep learning approaches in breast cancer histopathology. *npj Breast Cancer,* vol. 9**,** no.1, p. 21.

Mcleish, E., Slater, N., Mastaglia, F. L., Needham, M. and Coudert, J. D. (2024). From data to diagnosis: how machine learning is revolutionizing biomarker discovery in idiopathic inflammatory myopathies. *Briefings in Bioinformatics,* vol. 25**,** no.1, p. bbad514.

Raza, M. L., Hassan, S. T., Jamil, S., Hyder, N., Batool, K., Walji, S. and Abbas, M. K. (2025). Advancements in deep learning for early diagnosis of Alzheimer's disease using multimodal neuroimaging: challenges and future directions. *Frontiers in Neuroinformatics,* vol. 19, p. 1557177.

Sadr, H., Nazari, M., Khodaverdian, Z., Farzan, R., Yousefzadeh-Chabok, S., Ashoobi, M. T., Hemmati, H., Hendi, A., Ashraf, A., Pedram, M. M., Hasannejad-Bibalan, M. and Yamaghani, M. R. (2025). Unveiling the potential of artificial intelligence in revolutionizing disease diagnosis and prediction: a comprehensive review of machine learning and deep learning approaches. *European Journal of Medical Research,* vol. 30**,** no.1, p. 418.

Wang, X. and Ji, J. (2025). Explainable machine learning framework for biomarker discovery by combining biological age and frailty prediction. *Scientific Reports,* vol. 15**,** no.1, p. 13924.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights