DOI: 10.53469/jtpes.2024.04(02).07

Utilizing AI-Enhanced Multi-Omics Integration for Predictive Modeling of Disease Susceptibility in Functional Phenotypes

Yanlin Zhou[1, *], Xinyu Shen[2], Zheng He[3], Huiying Weng[4], Wangmei Chen[5]

[1] Computer Science Johns Hopkins University, Baltimore, MD 21218
[2] Biostatistics Columbia University, USA
[3] Applied Analytics, Columbia University, NY, USA
[4] Master of Science in Information Studies, Trine University, Phoenix AZ, USA
[5] Computer Science (software technology), The national university of Malaysia, Malaysia
**Correspondence Author, popojoyzho@gmail.com*

Abstract: ***With the continuous development of machine learning technology, the scientific research of biomedical materials is gradually shifting to a data-driven direction. The rise of this trend stems from the widespread use of Bio sequencing technology, which provides entirely new methods and insights for testing and evaluating the biological function of biomedical materials. The performance and performance of biomedical materials have a wide range of applications in medical applications, drug delivery, biosensors and other fields, so it is important to further optimize them. However, with the accumulation and increasing complexity of data, there is a need for more intelligent and efficient ways to process and analyze this heterogeneous scientific data. Therefore, the establishment of an open, shared infrastructure for storing heterogeneous scientific data from different research fields will be the cornerstone of cross-disciplinary joint analysis. This infrastructure will not only accelerate the collection and integration of data, but will also provide opportunities for collaboration and innovation across disciplines. This paper highlights a new trend in biomedical materials research, namely a data-driven approach, and the key role of Bio sequencing technology in this process. At the same time, we call for the establishment of an open data storage and sharing platform to promote multidisciplinary cooperation, accelerate the optimization and innovation of biomedical materials, and open up broader prospects for future biomedical applications. This effort is expected to push scientific research in the medical field to new heights, providing safer and more effective treatments and medical programs for patients.***

Keywords: Biomedical Materials; Data-Driven; Biological Sequencing Technologies; Interdisciplinary Collaborative Analysis Logistics.

1. INTRODUCTION

With the rise of artificial intelligence technology boom again, the combination of medical field and AI technology is considered to be the most potential for development. It can be seen that the accumulation of artificial intelligence over the years, and the in-depth development of face speech recognition, deep learning and other fields have made AI technology continue to make breakthroughs in the medical field. Let's take a look at the development prospects of AI technology in the medical field. Among them, AI technology has played a great role in assisting medical data processing, which needs to extract clinical information and master the structure of data. In terms of extracting information, AI technology is needed to digitize and structure traditional unstructured text medical records, and then transform them into structured data that can be analyzed and processed. If AI technology is applied, the risk of serious complications can be predicted in advance, and then appropriate treatment can be given before chemotherapy to reduce the risk of serious complications. In addition, in the medical field, only by mastering more analyzable data structures can we make more scientific research results, which is also a prerequisite for the development of the medical field. Therefore, the core value of artificial intelligence for medical data processing is self-evident.

The second is the integration of AI and multi-omics and biological analysis. In recent years, metabolomics with small molecule analysis of metabolites as the core has been paid more and more attention on the basis of traditional small molecule quantitative analysis. Because metabolites are the basic active substances in living organisms, coupled with the repeatability and quantification of analytical methods, they are considered to be one of the best ways to describe the state of living organisms. The data that each analysis method can provide is limited, and the in-depth study of life requires us to collect all knowledge and comprehensive data. From the perspective of omics, it is to use the "multi-omics" method, using the cross and complement between the omics, as far as possible more systematically at the molecular level, in order to solve complex scientific problems.

Therefore, disease susceptibility studies are essential for predicting an individual's risk for a particular disease. It contributes to personalized medicine, early diagnosis, effective prevention and a deeper understanding of biology. Therefore, in this paper, AI can analyze large-scale genomic data, identify biomarkers, analyze medical images, and integrate multiple data sources to build predictive models, thereby helping to evaluate disease susceptibility and improve support for medical decisions. These applications can help improve disease management and prevention.

2. RELATED WORK

With the continuous emergence of new omics technologies, omics research has accelerated the development of quantitative, high-throughput direction, through the integration of multiple omics data analysis, has become a new direction for scientists to explore the mechanism of life. This chapter introduces methods for integrating multi-omics data (genomics, transcriptomics, proteomics, metabolomics, imagomics).

2.1 Geonomics

Genomics and Genomics gene: The basic unit of heredity, the nucleic acid segment that encodes RNA or polypeptide chains. genome: The complete set of haploid genetic material contained in a cell or organism. Genomic DNA: structural genes that encode proteins, regulatory sequences that replicate transcription, and regions whose functions are not yet known. The characteristics of the genome: (1) The size and complexity of the genome varies from organism to organism. (2) The more evolved organisms are, the more complex their genomes are. genomics: Proposed in 1986, Genomics is defined as the study of the structural composition, temporal expression patterns, and function of the genome and the provision of evolutionary information about biological species and their cellular functions. Genomics consists of three distinct subfields: structural genomics, functional genomics, and comparative genomics.

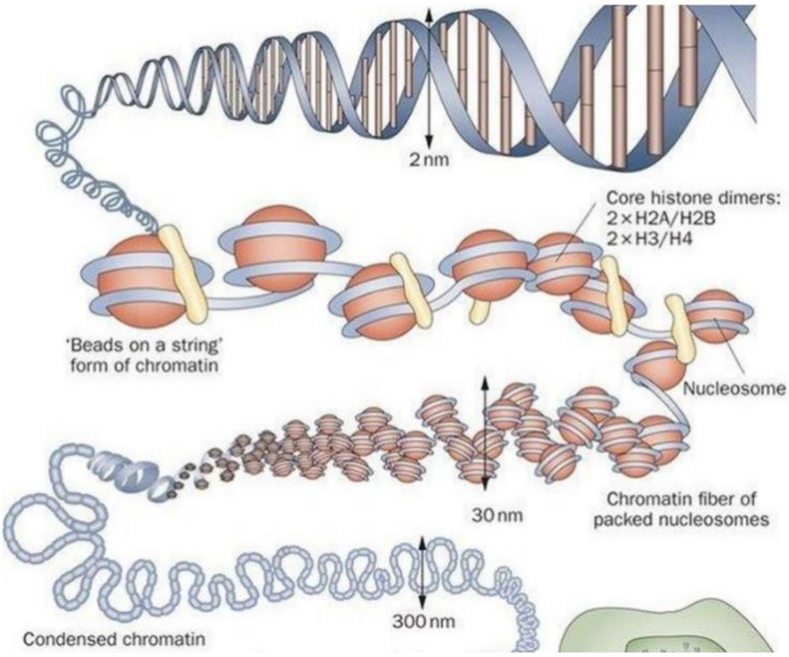

Figure 1: Genomic structure

Human Genome Project HGP human five model organisms: E. coli, fruit fly, mouse, nematode, yeast. The initial goal of the HGP was to spend $3 billion over 15 years (1990-2005) to complete 3 billion nucleotide sequence analyses of 24 human chromosomes. The ultimate goal of HGP is to decode life, understand life, understand the causes of differences between species and individuals, understand the mechanism of disease and life phenomena such as longevity and aging, and provide scientific basis for the diagnosis and treatment of diseases. The research goal of HGP is to make high-resolution genetic map of the human genome; Mapping the various physical maps of the human and certain model genomes; Determination of the complete DNA sequence of humans and certain model organisms; Collect, store, disseminate and analyse the resulting data and develop a range of new technologies for this purpose.

2.2 Transcriptomics

Transcriptomics is an important means of functional genome research, which includes complete transcripts of mRNA and non-coding RNA. The expression of the same gene is often different in different tissues and at different times. Transcriptomics can study the gene expression of specific cells, tissues or organs in different growth and development stages or under different physiological conditions at the RNA level and dig out key differential genes with specific biological functions. Prediction of lncrnas with regulatory functions and mirnas with negative regulatory functions; The regulatory mechanism of circRNA competitive endogenous RNA (ceRNA) was revealed. And complex networks of mutual regulation. RNA-seq technology based on high-throughput sequencing is the main means of transcriptomic research at present. Due to its advantages of high sensitivity, low noise and wide detection range, it is widely used in the mining of functional genes and the study of molecular genetic network regulation mechanism in livestock and poultry.

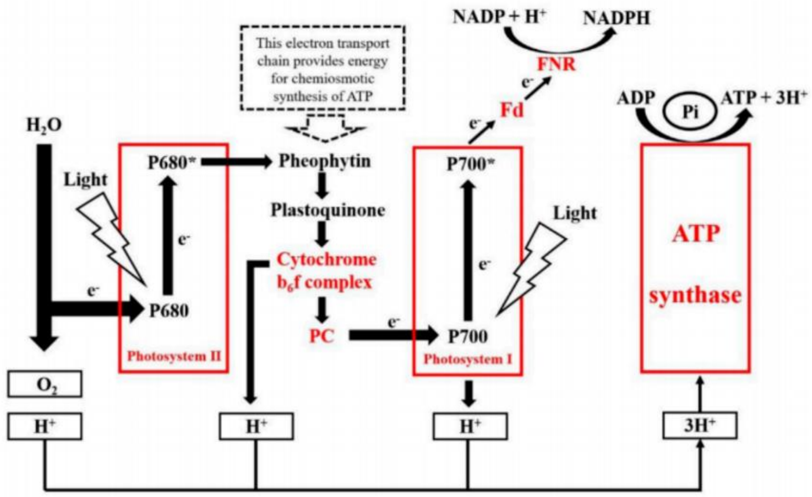

Figure 2: Transcriptomics enables differences in photosynthesis.

The transcriptomes in the previous joint analysis were all reference transcriptomes, and the non-reference transcriptome can also be combined and the relationship between genes and metabolites can be explored. in the article "Integrated analysis of the transcriptome and metabolome in young and mature leaves of Ginkgo biloba In L., researchers analyzed young and mature leaves of Ginkgo biloba by using metabolome and transcriptome, among which the transcriptome was a reference free transcriptome. During correlation analysis of differential metabolites, it was found that most metabolites were positively correlated with each other, and only serine was negatively correlated with other metabolites. The correlation analysis of differential metabolites and differential genes also found a negative correlation between serine and differential genes. In addition, four genes were also found to have a high correlation with metabolites, which can be used as follow-up research.

2.3 Proteomics

The term "proteomics" first appeared in 1995 and is defined as the large-scale characterization of all the proteins of a cell line, tissue, or organism. There are currently two definitions of proteomics, the first being a more classical definition that limits large-scale analysis of gene products to studies involving only proteins. The second and more inclusive definition combines protein studies with analyses with genetic readout, such as mRNA analysis, genomics, and yeast two-hybrid analysis.

However, the goal of proteomics remains the same, namely to obtain a more comprehensive and integrated view of biology by studying all the proteins of a cell rather than each protein individually.

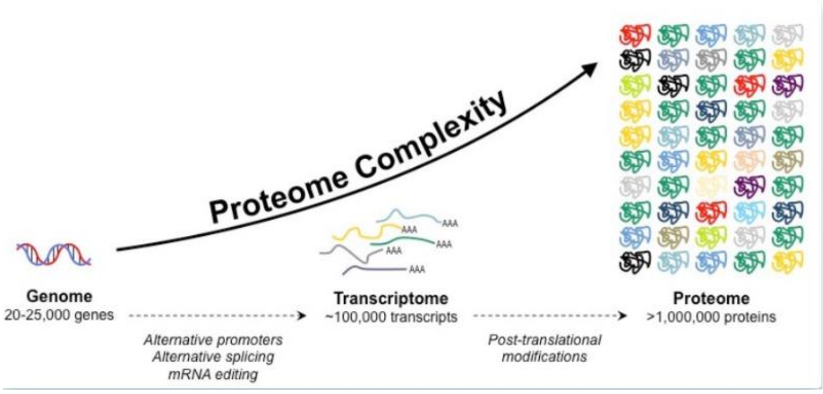

Figure 3: Protein structure diagram

A proteome is a group of proteins produced in an organism, system, or biological environment. For example, we can refer to the proteome of a species (such as a person) or an organ (such as the liver). The proteome is not constant; It varies from cell to cell and changes over time. To some extent, the proteome mirrors the underlying transcriptome, but protein activity is regulated by many factors in addition to productivity.

The combination of multiple omics data types, such as genomics, transcriptomics, proteomics, etc. with artificial intelligence has become an important trend in the biomedical field. This combination allows AI technologies such as machine learning and deep learning to analyze, integrate, and interpret this complex biological data, enabling more precise disease diagnosis, personalized treatment, and new drug development. The advantage of AI is that it can process large-scale multi-omics data, identify potential biomarkers, discover new therapeutic targets, and provide insight, accelerating the progress of biomedical research and promising to revolutionize the practice of medicine in the future.

3. METHODOLOGY

The aim of this study is to improve the accurate prediction of key proteins. Using artificial intelligence technology, we combine the topological structure of the protein interaction network (PPI), annotated gene ontology data and subcellular localisation information, while cleverly incorporating protein domain information. We propose a completely new algorithm, called TGSD, to identify these key proteins.

3.1 Algorithm Overview

The core idea of the TGSD algorithm in this experiment is to comprehensively consider multiple sources of information, including edge clustering coefficient, gene ontology annotation information and subcellular localisation data, to quantify the criticality of proteins. We also introduce protein domain information to reduce the impact of noise on the

Key design of protein domains Protein domains are the building blocks of proteins, A protein usually contains one or more domain information, in order to measure the importance of the domain in the protein, we combined the known key protein information to statistically analyze the importance of the protein domain, and defined the key value of the i-protein domain (PDV) as:

$$PDV(i) = VK(i) * VU(i) / \mathrm{Max}(PDV) \tag{1}$$

$$\begin{cases} VK(i) = \dfrac{|DP(i) \cap KKP| + 1}{|N(key)|} \Big/ \mathrm{Max}(VK) \\ VU(i) = \dfrac{|N - N(key)|}{|DP(i) \cap (P - KKP)| + 1} \Big/ \mathrm{Max}(VU) \end{cases} \tag{2}$$

Where, PDV(i) represents the fraction of protein domain i, the protein domain with a higher score is more critical, N represents the total number of proteins in the network, and N(hey) represents the number of key proteins in the network. DP(i) is made up of all proteins containing protein domain i KKP is the set made up of known key proteins and P is the set made up of all proteins with known protein domain data.

3.2 PPI Network Data

Protein interactions in yeast are the most extensively studied of all species, and a large amount of key protein data information has been accumulated for experimental verification. Therefore, four yeast PPI datasets, YDIP, DIP-PPI, Krogan and Krogan-extended, were selected for the experiment in this paper. After deleting the isolated nodes and repeated interaction relationship data from the original data, the detailed information of the dataset was shown in Table 1:

Table 1: Details of the data set

Data set	protein	interaction	Key protein	density
YDIP	5093	24743	1167	0.0019
DIP PPI	4928	17201	1150	0.0014
Krogan	2708	7123	786	0.0019
Krogan Extended	3672	14317	929	0.0021

The protein domain data used in this experiment were downloaded from the PFAM database and pre-processed according to the method proposed by Yang Zengguang et al. The pre-processed dataset contained 3630 proteins, including information data of 1107 protein domains, making up 4936 proteins. Gene Ontology Annotation data were downloaded from the Yeast Gene Ontology Annotation Database (10 September 2020 version), and subcellular localisation data and key protein data were obtained from the literature.

3.3 Experimental Result

In order to evaluate the effectiveness of the newly proposed algorithm, TGSD, TGSD and 7 representative algorithms (DC, BC, NC, Pec, WDC, etc.) are calculated respectively. LBCC and TEGS ranked the key values of proteins in 4 groups of yeast test data sets, and then ranked the key values of proteins in order from the largest to the smallest, and considered that the higher the ranking protein, the higher the probability of being the key protein: the eight methods were counted before the ranking The number of correct key proteins in N proteins, recognition accuracy, precision, etc., and the recognition effect of different methods were compared. Comparing the TGSD algorithm with other algorithms for predicting the correct number of key proteins in Figure 3, the TGSD algorithm is shown in Figure 3 comparison algorithms (DC, BC, NC, Pec, WDC, LBCC and TEGS) in the first 100, 200, ... Identify the correct number of key proteins in 600 proteins.

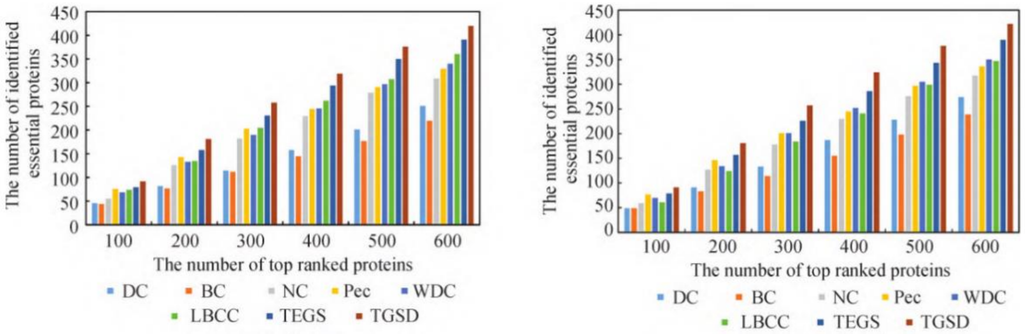

Figure 4: The TGSD algorithm and 7 other algorithms are sorting.

The numerical simulation results show that compared with the traditional DC, BC, NC, Pec, WDC, LBCC, TEGS and other algorithms, the TGSD algorithm has achieved a significant improvement in the prediction effect of critical protein. This research provides a powerful tool for the field of protein function research and bioinformatics, which is expected to lead us to a deeper understanding of the roles and regulatory mechanisms of key proteins in biological systems.

4. CONCLUSION

This study explores the shift of biomedical materials field towards data-driven direction with the development of machine learning technologies and highlights the key role of biosequencing technologies in this trend. With the wide application of biosequencing technology, this study proposes a new algorithm TGSD to identify key proteins, which has a wide application prospect in medical applications, drug delivery, biosensors and other fields. However, with the increasing accumulation and complexity of data, more intelligent and efficient methods of data processing and analysis are needed, hence the call for an open, shared multidisciplinary infrastructure to store heterogeneous scientific data from different research fields to accelerate multidisciplinary collaboration and innovation.

The integration of AI with multi-omics data has great potential, especially in the biomedical field. This combination enables the processing and analysis of large-scale biological data, including genomics, transcriptomics and proteomics, with the help of AI technologies such as machine learning and deep learning, enabling more accurate disease diagnosis, personalised treatment and new drug discovery. The power of AI lies in its ability to process multi-omics data, identify biomarkers, discover therapeutic targets and provide insights that accelerate the progress of biomedical research. In the future, we can expect broader applications of AI in disease susceptibility prediction, biomarker identification, genomics, transcriptomics and proteomics to support improved disease management and prevention.

Future research directions and potential application areas include further optimisation of AI algorithms to improve the prediction accuracy of key proteins and deepen the understanding of the roles and regulatory mechanisms of key proteins in biological systems. In addition, the integration of AI and multi-omics data bioanalysis can also be used for early diagnosis of diseases, drug screening, clinical decision support and other aspects, which is expected to promote the development of personalised medicine and precision medicine. At the same time, the development of open data storage and sharing platforms and the promotion of interdisciplinary cooperation will be an important direction in the future to better utilise multi-omics data and AI technology to solve complex problems in the biomedical field and provide better medical services to patients.

Acknowledgments

This study was supported by Dr. Shuqian Du and his 'Progress in the Application of Artificial Intelligence in Ultrasound Diagnosis of Breast Inspiration for the ideas and methods proposed in the article Cancer'. We would like to thank Dr. Shuqian Du and his team for providing us with valuable ideas and references. in addition, we would like to thank the journal 'Frontiers in Computing and Intelligent Systems' for providing us with a platform for communication in this area of research. This article provides us with insights and knowledge that will help move our research forward.

REFERENCES

[1] Zheng, Jiajian, et al. “The Credit Card Anti-Fraud Detection Model in the Context of Dynamic Integration Selection Algorithm”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 119-22, https://doi.org/10.54097/a5jafgdv.

[2] Qian, Jili, et al. “Analysis and Diagnosis of Hemolytic Specimens by AU5800 Biochemical Analyzer Combined with AI Technology”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 100-3, https://doi.org/10.54097/qoseeQ5N.

[3] Song, Tianbo, et al. “Development of Machine Learning and Artificial Intelligence in Toxic Pathology”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 137-41, https://doi.org/10.54097/Be1ExjZa.

[4] Du, Shuqian, et al. “Application of HPV-16 in Liquid-Based Thin Layer Cytology of Host Genetic Lesions Based on AI Diagnostic Technology Presentation of Liquid”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 1-6, doi:10.53469/jtpes.2023.03(12).01.

[5] “Based on Intelligent Advertising Recommendation and Abnormal Advertising Monitoring System in the Field of Machine Learning”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 17-23, https://doi.org/10.62051/ijcsit.v1n1.03.

[6] Yu, Liqiang, et al. “Research on Machine Learning With Algorithms and Development”. Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 7-14, doi:10.53469/jtpes.2023.03(12).02.

[7] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[8] Yu, L., Liu, B., Lin, Q., Zhao, X., & Che, C. (2024). Semantic Similarity Matching for Patent Documents Using Ensemble BERT-related Model and Novel Text Processing Method. arXiv preprint arXiv:2401.06782.

[9] Huang, J., Zhao, X., Che, C., Lin, Q., & Liu, B. (2024). Enhancing Essay Scoring with Adversarial Weights Perturbation and Metric-specific AttentionPooling. arXiv preprint arXiv:2401.05433.

[10] Tianbo, Song, Hu Weijun, Cai Jiangfeng, Liu Weijia, Yuan Quan, and He Kun. "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition." In 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), pp. 834-837. IEEE, 2023.DOI: 10.1109/mce.2022.3206678

[11] Liu, B., Zhao, X., Hu, H., Lin, Q., & Huang, J. (2023). Detection of Esophageal Cancer Lesions Based on CBAM Faster R-CNN. Journal of Theory and Practice of Engineering Science, 3(12), 36–42. https://doi.org/10.53469/jtpes.2023.03(12).06

[12] Liu, Bo, et al. "Integration and Performance Analysis of Artificial Intelligence and Computer Vision Based on Deep Learning Algorithms." arXiv preprint arXiv:2312.12872 (2023).

[13] “Implementation of Computer Vision Technology Based on Artificial Intelligence for Medical Image Analysis”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 69-76, https://doi.org/10.62051/ijcsit.v1n1.10.

[14] “Enhancing Computer Digital Signal Processing through the Utilization of RNN Sequence Algorithms”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 60-68, https://doi.org/10.62051/ijcsit.v1n1.09.

[15] Dong, Xinqi, et al. “The Prediction Trend of Enterprise Financial Risk Based on Machine Learning ARIMA Model”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 65-71, doi:10.53469/jtpes.2024.04(01).09.

[16] Tan, Kai, et al. “Integrating Advanced Computer Vision and AI Algorithms for Autonomous Driving Systems”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 41-48, doi:10.53469/jtpes.2024.04(01).06.

[17] “A Deep Learning-Based Algorithm for Crop Disease Identification Positioning Using Computer Vision”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 85-92, https://doi.org/10.62051/ijcsit.v1n1.12.

[18] Wang, Sihao, et al. “Diabetes Risk Analysis Based on Machine Learning LASSO Regression Model”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 58-64, doi:10.53469/jtpes.2024.04(01).08.

[19] Wei, Kuo, et al. “Strategic Application of AI Intelligent Algorithm in Network Threat Detection and Defense”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Jan. 2024, pp. 49-57, doi:10.53469/jtpes.2024.04(01).07.

[20] S. Tianbo, H. Weijun, C. Jiangfeng, L. Weijia, Y. Quan and H. Kun, "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition," 2023 3rd International Conference on Consumer Electronics and Computer Engineering (ICCECE), Guangzhou, China, 2023, pp. 834-837, doi: 10.1109/ICCECE58074.2023.10135464.

[21] Zhang, Quan, et al. “Deep Learning Model Aids Breast Cancer Detection”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 1, Dec. 2023, pp. 99-102, https://doi.org/10.54097/fcis.v6i1.18.

[22] Jingyu Xu, Yifeng Jiang, Bin Yuan, Shulin Li, Tianbo Song,Automated Scoring of Clinical Patient Notes using Advanced NLP and Pseudo Labeling,arXiv preprint arXiv:2401.12994, 2024

[23] Xiaonan Xu, Bin Yuan, Yongyao Mo, Tianbo Song, Shulin Li, Curriculum Recommendations Using Transformer Base Model with InfoNCE Loss And Language Switching Method, arXiv preprint arXiv:2401.09699

[24] “A Deep Learning-Based Algorithm for Crop Disease Identification Positioning Using Computer Vision”. International Journal of Computer Science and Information Technology, vol. 1, no. 1, Dec. 2023, pp. 85-92, https://doi.org/10.62051/ijcsit.v1n1.12.