Journal of Theory and Practice of Engineering Science ISSN: 2790-1505
DOI: 10.53469/jtpes.2024.04(02).09

Application of the AlphaFold2 Protein Prediction Algorithm Based on Artificial Intelligence

Quan Zhang[1,*], Beichang Liu[2], Guoqing Cai[3], Jili Qian[4], Zhengyu Jin[5]

[1,2,3,4] Information Studies, Trine University, Phoenix AZ, USA
[5] Informatics, Univeristy of California, Irvine,CA,USA
*Correspondence Author, wayne168zhang@gmail.com

Abstract: ***As the expression products of genes and macromolecules in living organisms, proteins are the main material basis of life activities. They exist widely in various cells and have various functions such as catalysis, cell signaling and structural support, playing a key role in life activities and functional execution. At the same time, the study of protein can better grasp the life activities from the molecular level, and has important practical significance for disease management, new drug development and crop improvement. Due to advances in high-throughput sequencing technology, protein sequence data has grown exponentially. The protein function prediction problem can be seen as a multi-label binary classification problem by extracting the features of a given protein and mapping them to the protein function label space. A variety of data sources can be mined to obtain protein function prediction features, such as protein sequence, protein structure, protein family, protein interaction network, etc. The initial steps are classical sequence-based methods, such as BLAST, which calculate the similarity between protein sequences and transmit annotations between proteins whose similarity scores exceed a specific threshold. This method has great limitations for protein function prediction without sequence similarity. Therefore, this paper analyzes the development prospect of bioanalysis and artificial intelligence through the application status and realization path of AlphaFold2 protein prediction algorithm based on artificial intelligence.***

Keywords: Artificial Intelligence; Biological Analysis; Protein Prediction; Alphafold2.

1. INTRODUCTION

Protein is an important life substance, which has many functions, such as catalyzing intracellular chemical reactions, providing cell structural elements, resisting diseases in the form of antibodies, controlling gene activity and regulating gene expression. There is a correlation between protein structure and protein function, and specific protein structure corresponds to specific function. Therefore, predicting protein structure is conducive to predicting protein function and drug discovery. Protein, as the expression product of genes and macromolecules in organisms, is the main material basis of life activities and widely exists in various cells. It has multiple functions such as catalysis, cell signaling and structural support, and plays a key role in life activities and functional execution. At the same time, the study of protein can better grasp the life activities from the molecular level, and has important practical significance for disease management, new drug development and crop improvement. Due to advances in high-throughput sequencing technology, protein sequence data has grown exponentially.

The protein function prediction problem can be seen as a multi-label binary classification problem by extracting the features of a given protein and mapping them to the protein function label space. A variety of data sources can be mined to obtain protein function prediction features, such as protein sequence, protein structure, protein family, protein interaction network, etc. The most commonly used sources of information are protein sequences and interaction networks. Generally speaking, the research of protein function prediction can be divided into three stages. The initial steps are classical sequence-based methods, such as BLAST, which calculate the similarity between protein sequences and transmit annotations between proteins whose similarity scores exceed a specific threshold. This method has great limitations for protein function prediction without sequence similarity. The second stage is the machine learning method based on decision tree and support vector machine, which is represented by multi-source K-nearest neighbor algorithm. Then moving to the third phase of the deep learning model, the researchers proposed the DeepGOPlus model, which does not rely on embedding vectors of protein nodes in the protein-protein interaction network, but instead captures sequence similarity information through sequence alignment tools and extracts sequence features in combination with CNN to improve prediction performance. DeepGraphGO uses the family and domain information of the sequence to provide the initial features for the nodes, and then uses the graph convolutional network to obtain the structure information of the PPI network.

On this basis, PSPGO proposes a multi-species labeling and feature propagation model based on protein sequence similarity network and PPI network.

In this paper, with protein sequence information and protein structure information as input, sequence features were extracted by SeqVec pre-training model, and structural features were extracted by hierarchical graph pooling model based on self-attention mechanism. In order to maximize the protein structure information provided by AlphaFold2, pre-trained residual-level embedding in the protein structure network is performed via Node2vec, which is then used as the initial node feature of the pooled model.

2. RELATED WORK

With the development of deep learning, large-scale protein language model (PLM) has made great achievements in protein prediction tasks, such as protein 3D structure prediction and various functional prediction. AlphaFold2, a revolutionary AI protein model, achieved atomy-level prediction accuracy on the CASP14 protein structure prediction task, a result that could reshape structural biology. When it comes to proteins, though, structure is just the beginning. For the interpretation of protein function, such as unknown protein function annotation, mutation effect, protein engineering, folding stability and other studies have more practical significance.

The quest for precise robot positioning within logistics automation has been a focal point in recent research endeavors. A multitude of techniques and methodologies have been explored to address this critical challenge.

2.1 Amino Acid and Protein Structure

Amino acid structure, amino group, alpha-carbon atom, carboxyl group and Side Chain free radical (R-group).

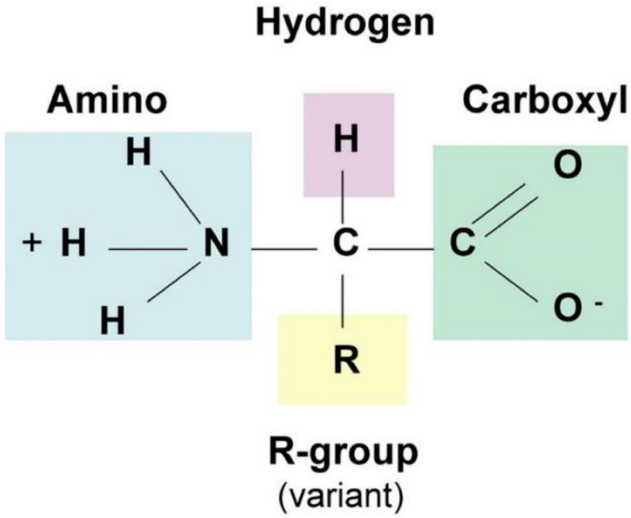

Figure 1: Amino acid structure diagram

Amino acids are joined together to form a Peptide. A Torsional Angle of the Peptide Plane is formed by rotating between adjacent Amino Acid Residue in a Peptide around the Peptide Bond.

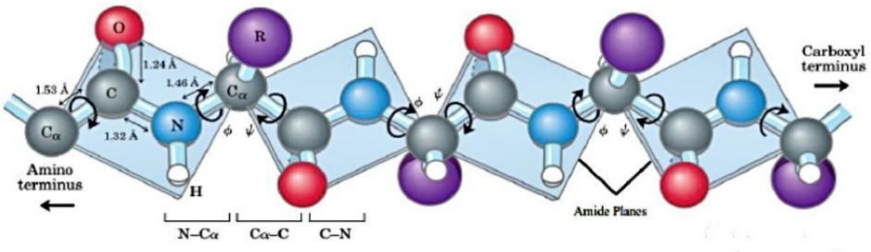

Figure 2: Amino acids in amino acids join together to form peptide structures.

Protein is the embodiment of life activities, and its structure determines its function. Proteins composed of linear amino acids need to fold into a specific spatial structure to have the corresponding physiological activity and

biological function. The analysis of the spatial structure of proteins is of great significance for understanding the function of proteins, the execution of functions, and the interaction between biological macromolecules. In order to understand the function of proteins more quickly, it is not only necessary to wait for the determination results of proteins, especially before the study of unknown proteins, through the prediction of protein structure has obvious advantages.

Proteins are the primary performers of life, so it is crucial to decipher the mechanisms behind their structural and functional properties. Among them, the protein sequence-structure-function relationship has made sequence-based machine learning methods very successful in structure and function prediction, which can infer protein structure and function from amino acid (AA) sequence. Large-scale protein language models with hundreds of millions of parameters have become the most mainstream approach for AI to predict protein function through self-supervised learning methods.

At the same time, AlphaFold was trained on 3D protein structures in the Protein Database (PDB) and could eventually output 3D protein structures that were as accurate as the experimental structures. Its multi-sequence alignment representation module Evoformer combines new deep learning mechanisms, PLM residual reconstruction tasks, and structural supervision in the form of histograms. Like MSA-Transformer, Evoformer uses a range of evolutionally relevant and aligned protein sequences as input, while PLMS such as ESM-1b and TAPE use only a single protein sequence.

2.2 AlphaFold2 Architecture principle

Principle of AlphaFold2 model gene.

1) The protein sequences of Homologues may vary greatly, but their structures may be consistent.
2) Two coevolved amino acid residues of the same protein sequence interact with each other.

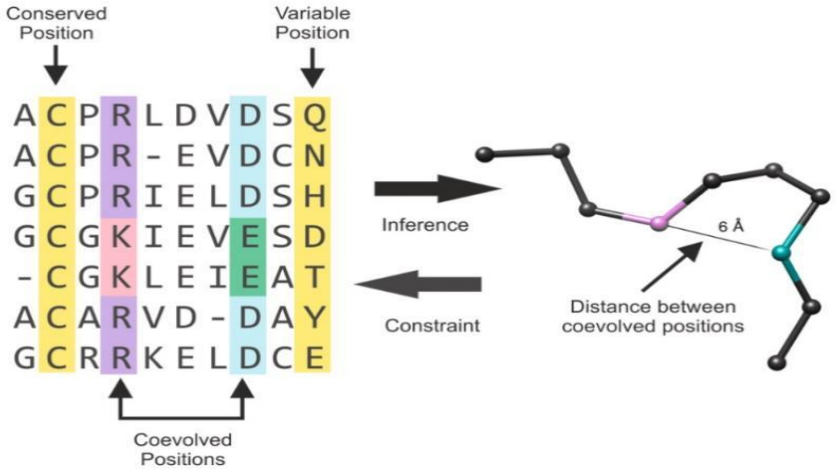

Figure 3: AlphaFold2 model architecture

The essence of AlphaFold2 is based on the protein sequence of homologous genes with the same protein structure and the relative position information of protein sequence atom pairs, and the Transformer mechanism is used to establish the characteristic corresponding relationship between amino acid residues and atoms in the case of sequence structure and three-dimensional geometry structure and protein structure. Multiple Sequence Alignment (MSA) information is formed by finding Homologues protein sequences from gene databases. The 3D Structure information of amino acid residues was found from the protein 3D Structure Template, and the Pair Representation of amino acid residues was established as the input data of the AlphaFold2 model. These data were extracted by Evoformer Encoder (Evoformer Encoder) to extract the characteristic information of protein sequence association of homologous genes and the characteristic information of distance between different amino acid residue pairs of the same protein.

The Structure Module calculates the three-dimensional coordinates of each atom in the protein sequence based on the feature information, compares them with the true value, and updates the weights of the AlphaFold2 model. The prediction is based on the three-dimensional coordinates of each atom, and the predicted protein structure is

displayed through three-dimensional image rendering. Therefore, an important step in the overall framework is Recycling, which can use the previous model output as input to improve prediction accuracy.

2.3 Computational Characteristics of AlphaFold2 Protein Structure Prediction

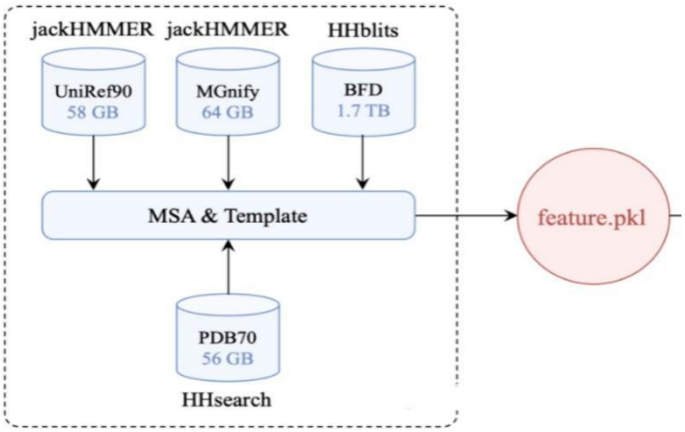

Figure 4: Sequence feature generation calculation

Calculation process:

Total input single protein sequence FASTA format (inference); The MSA (sequence-residue) was generated by a hidden Markov model search tool (jackHMMER/HHblits) on multiple genetic databases. Pairing information (residue-residuals) generated by the structure and sequence of the search; Template searched by HHsearch. The database search process involves intensive I/O reading and writing, and data is stored on high-speed SSDS. The accumulated data exceeds 2TB, which is time-consuming. The acceleration method improves the CPU computing speed.

In the process of implementation, multiple sequence alignment (MSA) is used to integrate protein Structure and biological information into deep learning algorithms, mainly including: neural network EvoFormer and Structure module. In EvoFormer, structure prediction is mainly completed by combining Graph networks and MSA. Alphafold2 uses Transformer structure, and Attention mechanism is used for information update of both MSA and residual-residue pairs. The update of the structure module uses the triangle rule, which simplifies the calculation complexity and improves the accuracy. The main job of the Structure Module is to convert the information obtained by EvoFormer into 3D protein structures. Recycling is used in the Evoformer and Structure module of the entire model, which is to add output to input for repeated refinement and information refinement.

2.4 The Alphafold2 Protein Predicts Advantage

(i) Evoformer, the main module of AlphaFold, can generate representations that are useful for both structural and functional prediction, such as two protein structure prediction tasks, two functional annotation tasks, and two mutation fitness landscape prediction tasks.
(ii) The vector representation of Evoformer's output is useful for both protein-level and residue level prediction tasks.
(iii) Evoformer is superior to ESM-1b and MSA-Transformer in structural prediction and stability prediction of novel small proteins, but inferior to ESM-1b in other functional prediction tasks. Compared to the ESM-1b and MSA-Transformer, it performs poorly on zero-sample fitness prediction tasks.
(iv) Evolutionarily aware PLMs only outperformed the non-evolved ESM-1b model for structural prediction tasks, but generally outperformed ESM-1b for most functional prediction tasks.
(v) MSA-Transformer and Evoformer are also very sensitive to the amount of MSA when predicting protein function. In addition, when using the MSA constructed by ESM-1b as input, the performance of the model is comparable to that generated by Jackhmmer or HHblits, but the speed is greatly improved. This study also proposes a deep learning approach to generate MSA quickly and accurately.

3. METHODOLOGY

The prediction of protein function in this experiment included metal ion binding and antibiotic resistance. And protein stability prediction, protein fluorescence prediction, and zero sample transfer learning mutation fitness landscape prediction. The experiment used ESM-1b, MSA-Transformer and Alphafold2 for a range of tasks, including protein structure prediction such as secondary structure, contact graph prediction.

3.1 Secondary Structure Prediction

In prediction experiments, base-level sequence-to-sequence tasks, where protein sequences x = {x_1,x_2,... .. Each residue x_i of,x_L} maps to the label y_i corresponding to eight secondary structure tasks y_i∈ {G, H,... C} one. Secondary structure prediction checks the extent to which PLM learns local structure.

For a given protein structure, two residues are considered to be in contact if they are within 8A of C_ beta carbon. We evaluate amino acids that are more than 6 locations apart in the primary structure. The PrecisionL measurement results are used, which represent the accuracy of Top-L amino acid pairs with the highest predicted contact probability. L is the length of the protein sequence.

3.2 Classification of Predictive Evaluation Tasks

1) Metal ion Binding (MIB): This is a binary classification task in which PLM is used to determine whether a metal ion binding site is present in a protein.
2) Antibiotic resistance (ABR): This is a multi-class classification task, PLM needs to correctly identify the antibiotic class of protein degradation. We built a dataset from CARD that contained 19 different antibiotic types.

And three fitness prediction tasks. Unlike functional annotation predictions, the protein sequences in this task are all from the same wild type with a small number of mutant residues.

1) Stability: This is a protein-level regression task that predicts that a protein can maintain its folded protease concentration.
2) Fluorescence: This is also a protein-level regression task that predicts the logarithmic fluorescence intensity of a protein sequence.
3) Zero sample mutation effect prediction: This is a protein-level prediction task that establishes a relationship with the protein fitness landscape by comparing the difference between the likelihood probability assigned to the mutant residue and the likelihood assigned to the wild type. This subtask considers only single-point mutation data.

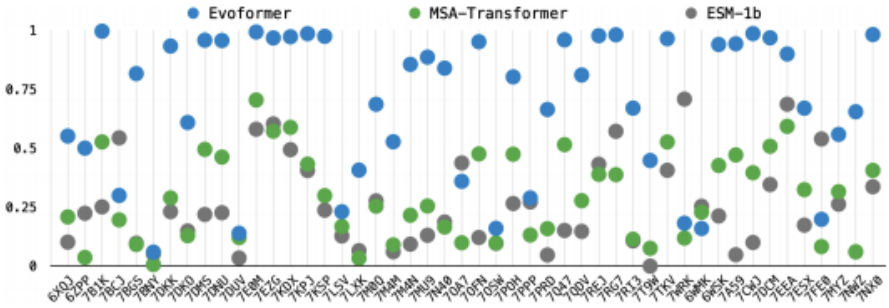

Figure 4: Prediction of new protein contact map

3.3 Protein Function Prediction Results

(1) Protein function annotation prediction:

Table 1: Protein function prediction results table

	Pre-train		Scratch	
Model	MIB	ABR	MIB	ABR
ESM-1b	**0.840**	**0.979**	0.628	0.945
MSA-Transformer	0.715	0.961	0.640	0.932
Evoformer	0.794	0.979	0.645	0.920

(2) Landscape prediction of protein mutation adaptability:

Table 2: Table of landscape prediction results of protein mutation fitness

	Pre-train		Scratch	
Model	Fluorescence	Stability	Fluorescence	Stability
One-hot [28]	0.14	0.19	-	-
ResNet [28]	0.21	0.73	0.28	0.61
ESM-1b	**0.68**	0.76	0.68	0.59
MSA-Transformer	0.64	0.67	0.67	0.61
Evoformer	0.67	**0.79**	0.36	0.52

(3) The following figure shows the landscape prediction results of zero sample mutation fitness:

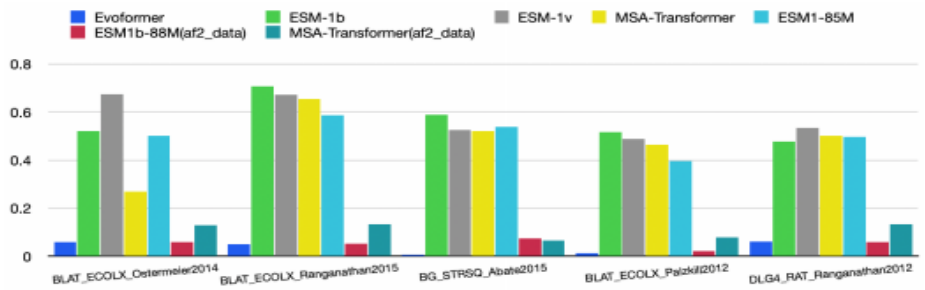

Figure 5: Zero sample mutation fitness score prediction graph

(4) The influence of the number of MSA on the model effect is tested on the two function prediction tasks of MIB and ABR. In the case of losing MSA, Evoformer and MSA-Transformer will produce worse function prediction results regardless of whether they are pre-trained:

Table 3: Table of the influence of MSA on the model results

Table 6: Impact of MSAs. 'Seq' denotes an individual sequence, i.e., no MSAs.

		SS		MIB		ABR	
Model	Pretrained	MSA	Seq	MSA	Seq	MSA	Seq
Evoformer	Yes	0.785	0.716	0.794	0.724	0.979	0.983
MSA-Transformer	Yes	0.748	0.631	0.715	0.707	0.961	0.908
Evoformer	No	0.614	0.624	0.645	0.632	0.920	0.875
MSA-Transformer	No	0.634	0.526	0.640	0.579	0.932	0.909

3.4 Experimental Conclusion

This study presents a deep learning approach aimed at predicting changes in protein solubility after mutation. This method is the first to use AlphaFold technology to predict the three-dimensional structure of proteins before mutation, which is a pioneering attempt in the field of predicting solubility changes after mutation. In addition, in order to make up for the deficiency of the data set, the researchers also used the pre-trained protein language model and the dissolved value prediction model for knowledge transfer, which improved the efficiency and accuracy of feature acquisition.

Compared with the traditional method, this new method shows significant advantages in computational efficiency and prediction performance. In particular, the use of features generated by pre-trained models not only improves the computational speed, but also surpasses other existing techniques in terms of prediction accuracy. In addition, the effectiveness of the knowledge transfer technique in predicting the solubility of protein mutations was experimentally verified.

Finally, the researchers used the DeepMutSol model they developed to predict changes in solubility on clinical pathogenicity related datasets. This prediction not only reveals the correlation between mutation solubility changes and pathogenicity, but also provides a powerful tool for disease pathway analysis and new drug discovery.

4. CONCLUSION

This study proposes a deep learning approach aimed at predicting changes in protein solubility after mutation. The method is the first to use AlphaFold technology to predict the three-dimensional structure of a protein before mutation, which is a pioneering attempt in the field. In addition, to make up for the deficiency of the data set, the researchers also used the pre-trained protein language model and the dissolved value prediction model for knowledge transfer, improving the efficiency and accuracy of feature acquisition. Compared with traditional methods, this new method shows significant advantages in computational efficiency and predictive performance. In particular, features generated using pre-trained models not only increase computational speed, but also surpass other existing techniques in terms of predictive accuracy. In addition, the effectiveness of knowledge transfer techniques in predicting the solubility of protein mutations has also been verified experimentally.

Artificial intelligence is playing an increasingly important role in protein prediction and bioinformation analysis. In particular, techniques that utilize deep learning and large-scale protein language models, such as AlphaFold2, provide unprecedented accuracy and efficiency for protein structure prediction and functional prediction. With these techniques, researchers are able to predict the structure and function of proteins more quickly, which has important implications for drug discovery, disease management and crop improvement.

Conclusion, AI is also able to make protein function prediction using a large amount of biological data, including data sources such as protein sequences, protein structures, protein families, and protein interaction networks. With deep learning models, features can be extracted from these data and mapped into the protein function label space, thus enabling prediction of protein function. The development of these technologies has brought new breakthroughs in the field of life science research and medicine, and made important contributions to human health and social well-being.

REFERENCES

[1] Bæk KT, Kepp KP. Assessment of AlphaFold2 for Human Proteins via Residue Solvent Exposure. J Chem Inf Model. 2022;62(14):3391-3400.
[2] “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.
[3] 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.
[4] 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.
[5] 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.
[6] 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
[7] “The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier”. Academic Journal of Science and Technology, vol. 8, no. 2, Dec. 2023, pp. 57-61, https://doi.org/10.54097/ajst.v8i2.14945
[8] Pan, Yiming, et al. “Application of Three-Dimensional Coding Network in Screening and Diagnosis of Cervical Precancerous Lesions”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 61-64, https://doi.org/10.54097/mi3VM0yB.
[9] 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
[10] 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).
[11] 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).
[12] 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.

[13] “The Application of Artificial Intelligence to The Bayesian Model Algorithm for Combining Genome Data”. Academic Journal of Science and Technology, vol. 8, no. 3, Dec. 2023, pp. 132-5, https://doi.org/10.54097/ykhccb53.
[14] 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.
[15] 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.
[16] Xin, Q., He, Y., Pan, Y., Wang, Y., & Du, S. (2023). The implementation of an AI-driven advertising push system based on a NLP algorithm. International Journal of Computer Science and Information Technology, 1(1), 30-37.0
[17] Pan, Yiming, et al. “Application of Three-Dimensional Coding Network in Screening and Diagnosis of Cervical Precancerous Lesions”. Frontiers in Computing and Intelligent Systems, vol. 6, no. 3, Jan. 2024, pp. 61-64, https://doi.org/10.54097/mi3VM0yB.
[18] “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.
[19] Chen, Wangmei, et al. “Applying Machine Learning Algorithm to Optimize Personalized Education Recommendation System”. Journal of Theory and Practice of Engineering Science, vol. 4, no. 01, Feb. 2024, pp. 101-8, doi:10.53469/jtpes.2024.04(01).14.
[20] K. Jin, Z. Z. Zhong and E. Y. Zhao, "Sustainable Digital Marketing Under Big Data: An AI Random Forest Model Approach," in IEEE Transactions on Engineering Management, vol. 71, pp. 3566-3579, 2024, doi: 10.1109/TEM.2023.3348991.
[21] “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.
[22] 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.
[23] “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.
[24] 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.