Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering

Zheng He*
Applied Analytics, Columbia University, NY, USA
zh2541@columbia.edu

Xinyu Shen
Biostatistics Columbia University, TX, Frisco USA
766761558@qq.com

Yanlin Zhou
Computer Science Johns Hopkins University, MD, USA
m24296170@gmail.com

Yong Wang
Information Technology, University of Aberdeen, Aberdeen, United Kingdom
13107712116@163.com

ABSTRACT

In the field of bioinformatics, the algorithm of biological gene sequence has always been one of the hot problems in scientific research. With the combination of biology and artificial intelligence technology, genetic algorithm data is increasing. At the same time, the emergence of a new generation of sequencing technology, the decrease in sequencing time and cost, and the high sequencing throughput have significantly increased the sequence data, showing an exponential growth trend, and there are still new biological gene sequence data found and recorded every day, and the speed of data generation is much faster than the speed of data processing, so the processing of large-scale DNA sequencing data needs more efficient methods. Therefore, AI bioinformatics engineering, especially genetic algorithms and K-means cluster analysis, has become an important tool in the field of bioinformatics statistics, with particular impact on the detection and diagnosis of single nucleotide polymorphisms (SNPS) associated with contact dermatitis. These advanced AI models efficiently process large amounts of genomic data, including SNP datasets, and have the ability to analyze patient genotype information. By doing so, they can identify SNPS that are strongly associated with contact dermatitis and establish meaningful associations between these genetic variants and the disease.The adoption of this personalized medicine approach not only addresses the specific needs of individual patients but also significantly enhances the success rate of treatment. Through the analysis of convolution algorithm and gene sequence in biological information engineering, this paper focuses on the relevant experiments of gene K-cluster analysis model, demonstrating the advantages and reference significance of gene algorithm under K-cluster analysis for current gene information statistics.

CCS CONCEPTS

• **Networks** → Network performance evaluation; Network performance analysis.

BIC 2024, January 26–28, 2024, Beijing, China

ACM ISBN 979-8-4007-1664-5/24/01
https://doi.org/10.1145/3665689.3665767

KEYWORDS

K_ mean clustering, Disease detection, Single nucleotide polymorphism, Bioinformation process

ACM Reference Format:
Zheng He*, Xinyu Shen, Yanlin Zhou, and Yong Wang. 2024. Application of K-means clustering based on artificial intelligence in gene statistics of biological information engineering. In *2024 4th International Conference on Bioinformatics and Intelligent Computing (BIC 2024), January 26–28, 2024, Beijing, China.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3665689.3665767

1 INTRODUCTION

With the development of big data, cloud computing, Internet of Things and other technologies, deep learning and artificial intelligence have been widely used in many fields, such as speech recognition, computer vision, medical detection and gene identification. In the case of the explosive growth of large-scale data, how to use artificial intelligence models or algorithms to detect certain diseases in genes and make rapid diagnosis has become one of the most concerned problems in the current medical community [1]. The construction of deep neural network to make intelligent judgment on the risk coefficient and contact tracing of personnel in the intelligent epidemic prevention and control system provides a new idea for exploring the search of patients and contacts in the epidemic prevention and control system [2-4].

Therefore, by analyzing artificial intelligence technology, deep learning and genetic testing and screening, combined with biological information statistics, this paper has a profound impact on the field of medicine and life science, consolidating the application of [5] K-means cluster analysis in biological gene sequence, and thus contributing to the realization of new breakthroughs in personalized medicine and disease prevention. In addition, in biological information statistics, the application of genetic algorithm and K-means cluster analysis can better help people understand and mine genetic data, further improve the efficiency and accuracy of gene-related research, and realize the ability to process large data sets and extract meaningful insights.

2 RELATED WORK

Cluster analysis is an unsupervised learning method that divides a bunch of unlabeled data into several classes by finding clusters with high similarity and high dissimilarity. Cluster analysis is the earliest partition-based clustering method proposed by Macqueen

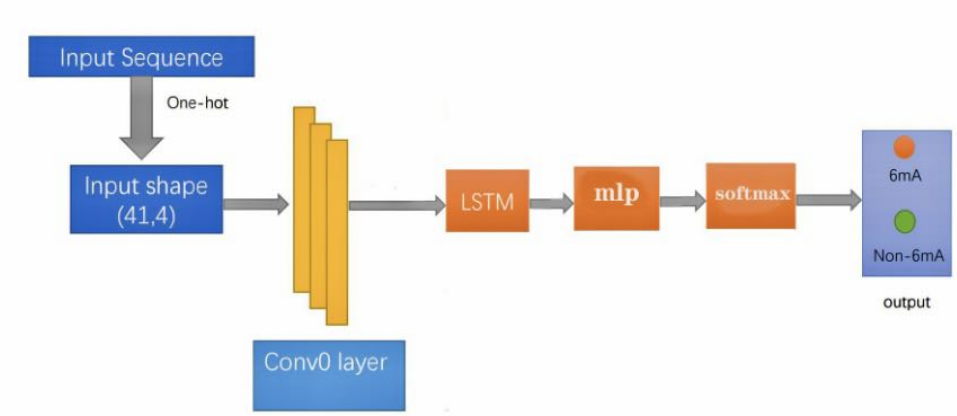

Figure 1: Convolutional neural network CNN predicts gene sequence model

-K-means clustering, which uses the average value of objects in the cluster to select the cluster center [6]. K-means clustering has fast convergence speed and simple algorithm, but its disadvantage is that random selection of K value leads to unstable clustering results. Therefore, many scholars have proposed different algorithms to improve the clustering center selection problem of K-means algorithm. Kaufman et al. proposed the K-medoids algorithm, which divides the class around the central point and selects the class closest to the central point several times. Huang et al [7]. proposed the K-modes clustering algorithm, which selects the class mode instead of the mean clustering through the mode of frequency, and uses simple matching dissimilarity instead of Euclidean distance to process cluster objects.

Under normal circumstances, the process of data object analysis must first divide the types of data objects, which can be classified or clustered. According to the clustering rule, data objects with high similarity are placed in the same class, while data objects with low similarity are abandoned. The result of the classification of multiple classes is called the clustering process [8]. Cluster analysis is also widely used in customer segmentation, such as real estate sales, Bioinformatics engineering, gene sequence analysis, medical diagnosis bank customer value assessment, air passenger value segmentation, etc. [9] In the application of genetic algorithm K_ mean cluster analysis in biological information statistics, there are several methods to quickly screen disease triggers by using the accurate and efficient data processing characteristics of large-scale artificial intelligence models:

2.1 Data collection and integration

In terms of the collection and integration of genetic data.Among these genetic data collection methods, the most common method models combined with artificial intelligence are machine learning models and deep learning, and machine learning models generally include decision trees, random forests, support vector machines, neural networks and other machine learning models for classification, regression and feature selection.

Figure 1 illustrates the application of a Convolutional Neural Network (CNN) in predicting gene sequences. These CNN models are utilized for analyzing correlations between genetic variations and particular traits or diseases. Deep learning, which encompasses the utilization of deep learning architectures like CNNs and Recurrent Neural Networks (RNNs), finds application in genomics for processing sequence data such as gene sequences or RNA sequences. This aids in the identification of potential functional elements or patterns in gene expression [10].

2.2 Feature extraction and engineering

Harness the power of AI models to feature extract and engineer the collected data to accurately reflect potential disease triggers. This can include extracting genotypes from genetic data, key indicators from clinical records, habits and risk factors from lifestyle data, and more. In general, for genome sequencing data, it is necessary to align and align the sequenced fragment with the reference genome. This can be done using sequence alignment algorithms such as Bowtie, BWA, or BLAST.

Figure 2 provides an illustration of relationships between alignment methods. The applications / corresponding computational restrictions shown are (green) short pairwise alignment / detailed edit model; (yellow) database search / divergent homology detection; (red) whole genome alignment / alignment of long sequences with structural rearrangements; and (blue) short read mapping / rapid alignment of massive numbers of short sequences. Although solely illustrative, methods with more similar data structures or algorithmic approaches are on closer branches. The BLASR method combines data structures from short read alignment with optimization methods from whole genome alignment.

2.3 Genetic algorithm K_ mean clustering

k-means cluster (KMCA) was performed using the Manhattan distance calculation. This approach enabled spectra grouping into classes based on their similarities and extraction of the average spectra reflecting the major biochemical classes of cells, originated from the whole cell cluster or the nucleus and cytoplasm fractions [9].

Representative Raman images of B lymphocytes (blue panel) and T cells (figure 3) obtained with 532 nm excitation are presented

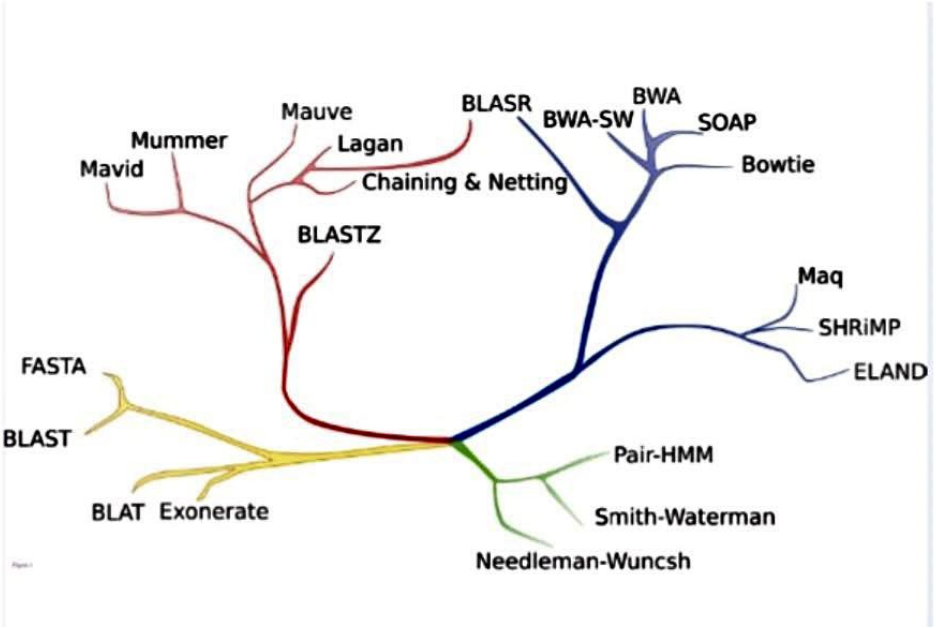

Figure 2: An illustration of relationships between alignment methods

in To visualize the size and shape of cells, the Raman bands corresponding to the C-H stretching vibrations (2800–3030 cm−1) were integrated.A single Raman spectrum contains a complete information about the structure of individual subcellular components, therefore, the visualization of individual cellular organelles is possible. The nucleus was imaged by integrating the Raman band in the range of 790–810 cm−1, which corresponds to DNA/RNA modes, whereas carotenoids were visualized by integrating their marker band in the spectral range of 1510–1530 cm−1, related to $\nu 1$ mode (Figure 3, b). Raman maps created by the integration of individual bands were compared with the color-coded KMCA maps representing major subcellular structures of interest, i.e. nucleus (blue), carotenoids (red), and cytoplasm (grey) from which the average spectra were extracted (Figure 3, d). In total, randomly selected 297 B cells and 464 T cells collected from 5 different donors were spectroscopically analyzed.It translated into the analysis of approximately 76,100 spectra of B lymphocytes and T lymphocytes in total, using 532 nm excitation and 25, 366 spectra using 633 nm excitation.While carotenoids class was observed for a significant proportion of studied T lymphocytes (Figure 3.a), it was found only in a few B cells. This observation led to the hypothesis that the presence of carotenoids may be a contributing factor in distinguishing B and T cells [11]. The K-means clustering algorithm relies on users to define the number of clusters (K) to find. Using a predefined number of clusters for all FCM samples is not possible due to intersample variability. We solved this problem by automatically choosing K based on a reasonable maximum.

3 METHODOLOGY

The clustering algorithm has been widely used in the analysis and processing of gene expression data. In this paper, K-means clustering algorithm is introduced into genetic algorithm, and a K-means clustering model based on genetic algorithm is discussed by combining the characteristics of gene microarray. The purpose is to improve the global nature of Litian genetic algorithm to improve the possibility of clustering algorithm to find global optimal. Experimental results prove. This algorithm can solve the cluster analysis problem of some gene expression data well.

3.1 K-means clustering model

K-means clustering is a classification clustering method. This algorithm is a very simple but commonly used method. Before cluster analysis, it is first assumed that n cluster objects can be divided into k classes and a representative of each class is determined, which is usually the center of gravity and the initial cluster point. Then, each cluster object is compared with these agglomeration points and reclassified according to the proximity between the cluster object and the agglomeration point, and the cluster object is classified into the category of the cluster center closest to it.

The algorithm steps are as follows:

(1) D selects k objects arbitrarily from n data objects as the initial clustering center

(2) Cycle the following flow 3 to 4 until each cluster no longer changes;

(3) According to the mean of each cluster object, calculate the distance between each object and these central objects, and re-divide the corresponding objects according to the minimum distance:

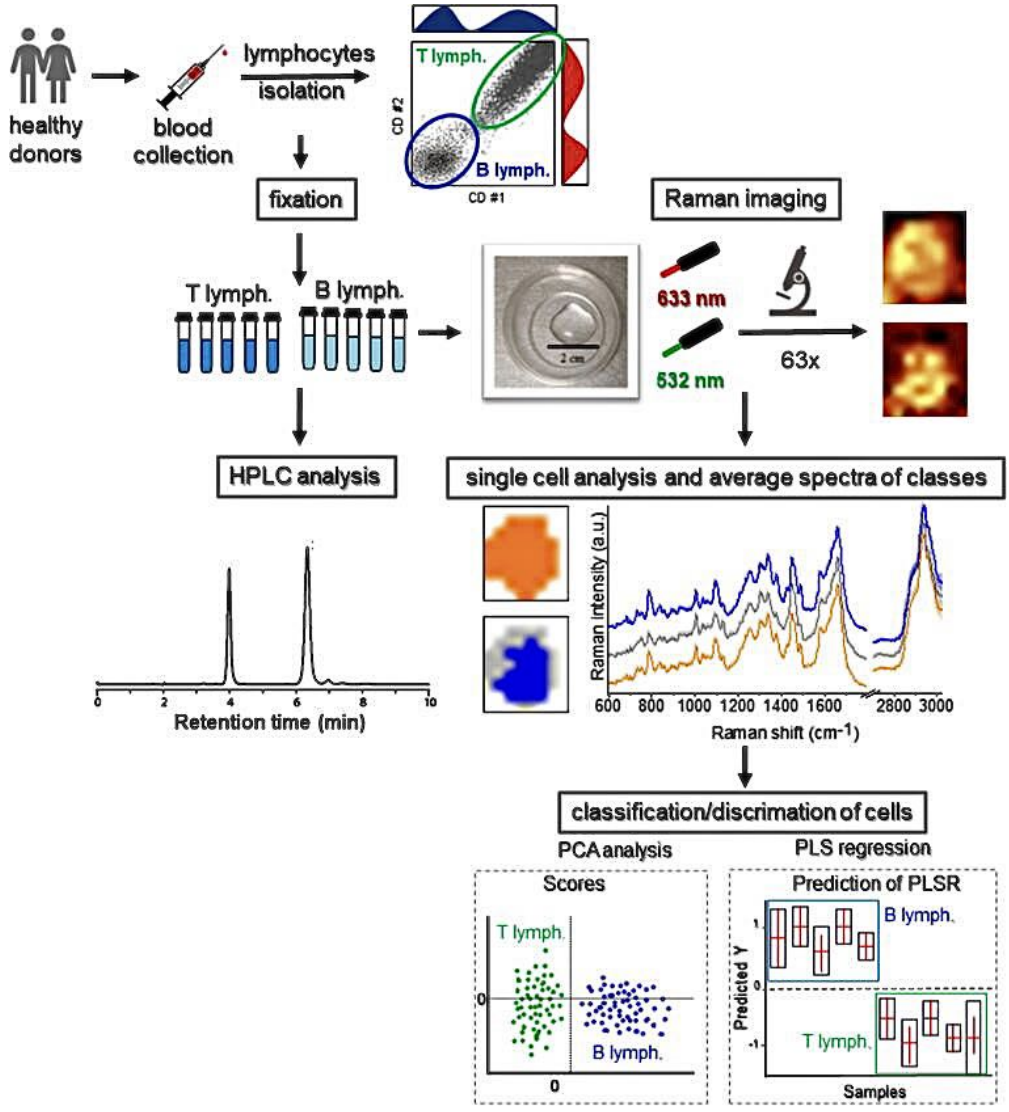

Figure 3: k-means cluster (KMCA) was performed using the Manhattan distance calculation model

(4) Recalculating the mean value of each cluster can generally be used as the standard measure function, which is defined as follows:

$$E = \sum_{i=1}^{k} \sum_{X \in ci} |x - Pi|2 \quad (1)$$

Where x is a point in the space representing the object, and Pi is the mean of the cluster ci.

The computational complexity of K-means algorithm is O(nkt), and nkt is the number of samples, the number of categories and the number of iterations respectively. Usually k<ntn, so K-means clustering can be applied to the case of large data volume [12]. This is one of its advantages. The key problem of the algorithm is how to select the initial cluster center.

3.2 K-means operation

Since the initial conditions are arbitrarily set, and the selection and mutation operations are probabilistic, the standard genetic algorithm takes a long time to converge. In order to improve the convergence speed, we introduce the k-means operation, and finally get the optimal solution S 'with the smallest intra-class squared error [13-14]. The K-means operation consists of the following two steps D to calculate the center of mass of each class in the individual Si; 2 Calculate the distance of each sample from each centroid and assign it to the nearest class to get S Retain the optimal individual.The most fit individuals in each generation population are stored in an array, and when the genetic process is over, these individuals are compared and the best individual is output as the result.

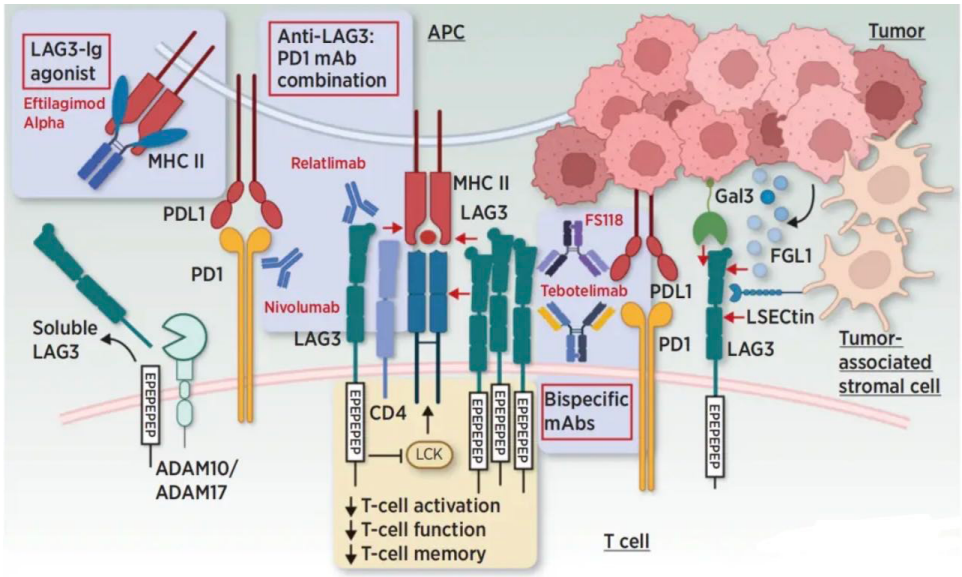

Figure 4: interaction and structural similarities between LAG3 and P14 T cells

3.3 Data source

P14 T cells were identified by manual gating of splenocytes stained with three sets of antibodies, and cells with the same stain were pooled as a population. Flow cytometry standard (FCS) files were imported into R from the Flowjo workspace.

Analysis of P14 Tcellsubsets was conducted in three steps. First, k-means clustering was performed by using kmeans function in R's Stats package to divide P14 T cells from each of the three populations into several subpopulations. Variables selected for cluster analysis of the three populations were: differentiation markers (CD62L, CD27, CD127 and KLRG1), transcription factors (EOMES, T-bet, Blimp-1, phospho-STAT3 and Bcl-6), and cytokines (IFN-γ, TNF-α, IL-2 and CCL3). The optimal number of clusters (k) was determined by plotting within groups sum of squares versus number of clusters, and k for the three populations (differentiation, transcription factor, cytokine) was set to 6, 6 and 5, respectively.

They are also structurally similar to P14T cells, with four IgSF-like domains (D1-14) of the extracellular immunoglobulin superfamily (figure 4), but the amino acid sequence homology of these two proteins is only about 20%. LAG-3 has a unique ring structure of about 30 amino acids in length in the distal D1 domain of the membrane, allowing LAG-3 to bind to MHC Class II molecules with a higher affinity than CD4. LAG-3 binds spatially to the TCR:CD3 complex to recruit signaling molecules and form immune synapses. The intracellular segment of LAG-3 has three distinct domains [15]: the potential serine phosphorylation site, the "KIEELE" element, and the "EP" repeating element. The highly conserved "KIEELE" element has been shown to mediate intracellular signal transduction and is necessary for LAG-3 to negatively regulate T cell function.

4 CONCLUSION

Bioinformatics is one of the major frontier fields of life science and natural science today, and will also be one of the core fields of natural science in the 21st century. Genomics and proteomics are two important aspects of bioinformatics. Specifically, it is the study of nucleic acids and proteins, their sequence, structure and function, and their evolutionary relationshipsThe experimental results show that the combination of genetic algorithm K-means cluster analysis and artificial intelligence (AI) shows strong potential in bioinformatics statistics and subcell population analysis. By adopting genetic algorithms to optimize the K-means clustering process, complex genomic data can be processed more efficiently and help identify patterns and characteristics of gene or subcell populations. This integration provides a highly flexible and adaptive approach that helps uncover hidden structures and associations in biological information, thereby advancing research in genomics and cell biology.

The K-measure values show that flow Means and [16] K-MEANS perform similarly, both on average and for individual samples (distributions of F-measures are shown in Supporting Information). In spite of using a more K-means-based automated gating framework. K-Means addresses all the issues that prevented the application of K-means to FCM data in the past. This makes k-Means a powerful tool for identification of cell populations as part of high throughput and accurate FCM data analysis.

ACKNOWLEDGMENTS

We sincerely thank Tianbo, Song, Hu and other authors for their contributions to the field of swarm intelligence. Their paper, "Bio-inspired Swarm Intelligence: a Flocking Project With Group Object Recognition not only provides us with a valuable research perspective, but also greatly promotes the development of group intelligence and target recognition technology. Many ideas and methodologies in this paper are inspired and guided by their research. We would like to thank them for their hard work in collating the data, building the model, and writing the paper:

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

REFERENCES

[1] Ripley B. tree: Classification and Regression Trees I . Wiley lnterdisciplinary Reviews Data Mining & Knowledge Discovery , 2018 1(1) : 14 23.

[2] Neshat M, Sepidnam G, Sargolzaei M, *et al.* Artificial fishswarm algorithm: a survey of the state of the art, hybridization, combinatorial and indicative applications [I] . Artificial Intelligence Review, 2014 , 42(4) : 965-997.

[3] Moustafa N,Slay J. UNSW-NB15: a comprehensive data setfor network intrusion detection systems (UNSW-NB15 net-work data set [J]) // Military Communications and Information Systems Conference. IEEE, 2015:16.Ambusaidi M, He X, Nanda P, *et al.* Building an intrusion detection system using a filter-based feature selection algorithm [W] . IEEE Transactions on Computers, 2016, 65.

[4] 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.

[5] Che, C., Liu, B., Li, S., Huang, J., & Hu, H. 2023. Deep learning for precise robot position prediction in logistics. Journal of Theory and Practice of Engineering Science, 3(10), 36-41.

[6] 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.

[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] Li, Linxiao, *et al.* "Zero-resource knowledge-grounded dialogue generation." Advances in Neural Information Processing Systems 33, 2020: 8475-8485.

[9] Ji S Y, Jeong B K, Choi S, *et al.* A multi-level intrusion detection method for abnormal network behaviors [J]. Journal of Network & Computer Applications, 2016, 62: 9-17.

[10] Prieto T , Alves J M , Posada D . NGS Analysis of Somatic Mutations in Cancer Genomes [M] //Big Data Analytics in Cenomics. Springer International Publishing, 2016.Larson D E , Harris C C , Chen K , *et al.* Somatic Sniper : identification of somatic point mutations in whole genome sequencing data . Bioinformatics, 2012, 28(3) : 311-317

[11] Moustafa N, Creech G, Slay J. Big data analytics for intru-sion detection system: statistical decision-making using finitedirichlet mixture models [M] / /Data Analytics and DecisionSupport for Cybersecurity. Springer, Cham, 2017: 127-156.

[12] Tian, Miao, *et al.* "The Application of Artificial Intelligence in Medical Diagnostics: A New Frontier." Academic Journal of Science and Technology 8.2, 2023: 57-61.

[13] Wan, Weixiang, *et al.* "Development and Evaluation of Intelligent Medical Decision Support Systems." Academic Journal of Science and Technology 8.2, 2023: 22-25.

[14] Pan, Linying, *et al.* "Research Progress of Diabetic Disease Prediction Model in Deep Learning." Journal of Theory and Practice of Engineering Science 3.12, 2023: 15-21.

[15] Liu, Yuxiang, *et al.* "Grasp and Inspection of Mechanical Parts based on Visual Image Recognition Technology." Journal of Theory and Practice of Engineering Science 3.12, 2023: 22-28.

[16] Zong, Yanqi, *et al.* "Improvements and Challenges in StarCraft II Macro-Management A Study on the MSC Dataset". Journal of Theory and Practice of Engineering Science, vol. 3, no. 12, Dec. 2023, pp. 29-35, doi:10.53469/jtpes. 2023 .03(12).05.