Proceedings of the 3rd International Conference on Mechatronics and Smart Systems
DOI: 10.54254/2755-2721/140/2025.21281

The Application of AI to Gene Classification

Yuanli Cheng[1,a,*], Silin Lin[2,b]

[1]College of Electronic and Information Engineering, Beibu Gulf University, Guangxi, China
[2]Beijing Huijia Private School, Beijing, China
a. 2204419733@stu.bbgu.edu.cn, b. 26linsilin@huijia.edu.cn
**corresponding author*

Abstract: Gene classification is important in disease identification, early diagnosis, and precise treatment. At the same time, traditional manual methods are limited by their time-consuming and low accuracy, and algorithms based on deep learning show outstanding potential. In this paper, the main research progress in the field of gene classification is reviewed, and the application characteristics and performance of deep learning models such as multi-layer perceptron (MLP), convolutional neural network (CNN), and recurrent neural network (RNN) are analyzed. In particular, the advantages of sci-kit-learn, Bidirectional Encoder Representations from Transformers (BERT), and DNASimCLR frameworks in gene classification are emphasized, and the importance of data preprocessing, feature extraction, and data set construction is discussed. At the same time, the limitations of current techniques are pointed out, including low training efficiency, sensitivity to outliers, and strong data dependence, and the possible direction of improving model performance through regularization techniques, pre-training models, data diversity enhancement, and other strategies are proposed. The research in this paper provides a reference for the optimization and popularization of gene classification technology and looks forward to its broad application in disease diagnosis and bioinformatics.

Keywords: Gene classification, neural network model, feature extraction and selection, data preprocessing, machine learning framework。

1. Introduction

In recent years, genetic classification technology has played an increasingly important role in disease identification and medical research. Analysis of gene expression data not only provides a clearer understanding of the molecular mechanisms of different diseases but also provides key support for early diagnosis and personalized treatment of diseases. However, traditional artificial gene classification methods have been unable to meet the current demand for massive data analysis due to their time-consuming, low efficiency, and limited accuracy. Especially when dealing with complex genetic data, the limitations of this method are more prominent. With the rapid development of artificial intelligence (AI) and deep learning technologies, these emerging tools have brought revolutionary changes to genetic classification. Deep learning methods not only significantly increase the speed of classification, but also enhance the precision of data processing. By automatically extracting features from large-scale gene expression data, AI can quickly complete sample classification and reveal its potential biological differences, thus injecting new vitality into medical

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/140/2025.21281

research and clinical practice. Especially in the early diagnosis of diseases and the formulation of treatment strategies, AI has provided strong technical support, while accelerating technological innovation and breakthroughs in the field of bioinformatics. In recent years, AI-based models have shown significant potential in gene classification. For example, the Bidirectional Encoder Representations from Transformers (BERT) model, as an advanced technique in the field of natural language processing, has been successfully applied to multiple aspects of gene classification. With the help of the BERT model, researchers can efficiently process the task of genetic mutation text classification, and effectively overcome the shortcomings of traditional manual classification, such as over-reliance on the professional judgment of pathologists, interference of redundant information, and inconsistencies in description. By training against challenges such as extreme text length, data bias, and high repeatability, the BERT model has performed well in terms of accuracy and adaptability and achieved remarkable research results [1]. In addition, other deep learning models have been explored for use in the analysis of genetic data for specific diseases. For example, the deep neural network (DNN) has been used to classify gene expression data for diseases such as leukemia, and it has shown good performance [2]. In the field of multi-task learning, researchers have explored the potential of genetic risk prediction for comorbidity neurodevelopmental disorders and found that these methods can effectively integrate features of different tasks, thereby improving the accuracy of the overall prediction [3]. These findings show that AI algorithms have broad application prospects in genetic classification and offer entirely new solutions for modern medicine in fields such as disease diagnosis and personalized treatment.

Based on this, this paper will deeply explore the specific applications of different machine learning methods in gene classification, describe the types of commonly used data sets, compare their advantages and disadvantages, explore potential ways to optimize these methods further and look forward to future development through the existing limitations, to promote the development of related research. The purpose of this review is to help researchers have a deeper understanding of the development status, challenges, and future development direction of gene classification technology, and to provide reference and guidance for the research and application in related fields.

2. Methods of Gene Classification

Gene classification usually includes multiple steps such as data preprocessing, feature extraction and selection, standardization, model training, and evaluation. First, data preprocessing will clean and preprocess the genetic data to extract meaningful features, such as nucleotide frequency, gene expression level, or protein sequence features. Then, a feature Selection method (e.g. Principal Component Analysis, Least Absolute Shrinkage, and Selection Operator) is used to reduce the redundant features and improve the generalization ability and performance of the model. To ensure the comparability of different features, genetic data is often standardized or normalized. Data segmentation divides the data into training sets, validation sets, and test sets to evaluate model performance. In the modeling process, gene classification usually uses a variety of neural network models, such as Multilayer Perceptron (MLP), Convolutional Neural Network (CNN), and Recurrent Neural Network (RNN). The input layer in the MLP will receive the gene expression profile and return the class probability of the gene expression sample through the output layer. CNN can convert the gene expression data into a two-dimensional image-like array with rows and columns, obtain the local spatial relationship between the input data, and then analyze and classify the gene expression. RNN can combine the input information with the previously stored information to generate a new output and can capture the correlation in the gene expression data sequence to achieve better classification. These three types of neural networks are commonly used in frameworks for various gene classification tasks, such as sci-kit-learn, BERT, and DNASimCLR. BERT can classify whether a gene is mutated by understanding the text data, and experiments show that the BERT+ abstract

DOI: 10.54254/2755-2721/140/2025.21281

truncation method makes the model perform best. DNASimCLR can extract complex features from different microbial gene sequences for pre-training, and then return classification results according to demand. scikit-learn framework was used to deal with the redundancy in genes by the method of feature selection, and classification models such as naive Bayes and support vector machine were used to classify aging genes, which could effectively improve the correctness of aging gene classification results [4]. He applied BERT's deep learning method to classify gene mutations based on textual evidence in annotated databases and solved some challenging problems, such as extreme text length, biased data presentation, and high repeatability [1]. Yang proposed the SimCLR framework, which can extract complex features from different microbial gene sequences and classify them based on convolutional neural network methods, showing excellent performance and strong adaptability in various tasks of analyzing biological sequence data [5].

3. Datasets

In genetic classification, data sets come from two main types for different purposes. One is the analysis of pathological sections. The main operation is to extract genetic information from pathological sections for sorting and training. Among them, microarray technology is a commonly used technology in pathological sections. Microarray data sets can be reserved for individual DNA sequences and genes at each point on the chip tool. These samples are hybridized and combined with microarray slides, and then lower and upper approximation properties of the genetic data are obtained by scanning the color of the spots on the slides to measure the expression of each gene. To divide the initial cluster into more clusters and screen out significant genes [2]. The other is to invoke existing data, which exists in a global database and can retrieve the literal or numerical genetic information that has already been formed. For example, in the text dataset of the database MSKCC, by analyzing the composition of the dataset, the construction algorithm summarizes the frequency of genes, gives them markers, extracts text entries shared by different samples, and classifies them [1].

4. Existing Limitations and Future Prospects

Neural networks may face efficiency problems when processing genetic data sets, especially when they are prone to falling into local minimums during training, resulting in the model failing to find the global optimal solution, resulting in sub-optimal classification results. In addition, the training process can become slow for huge gene data sets, especially if a large number of parameters need to be updated frequently. In response to this problem, there are several strategies that can be adopted to improve training efficiency. For example, regularization techniques can be used to give the machine some rules to avoid over-fitting and improve the generalization ability of the model. At the same time, the pre-trained model can be used to let the machine learn some general features first, and then fine-tune it for specific gene classification tasks, which can significantly improve the training efficiency. In terms of computational complexity, RNNS have higher computational complexity when dealing with long sequences, which can lead to slower processing, especially when each element of the sequence needs to be processed sequentially. To simplify the calculation process, the model structure can be simplified to reduce the complexity [6].

In the field of genomics, CNNS faces the dual challenges of data dependence and sensitivity to outliers. Due to the scarcity of genomic data, CNNS need a large amount of labeled data to learn effective feature representations, which limits their learning and generalization ability [7]. At the same time, the sensitivity of the MLP network to outliers or noisy data may affect the accuracy and robustness of the model. To improve the robustness and generalization ability of the model, it can first improve the adaptability of the machine by increasing the diversity of the data, reducing the need for large amounts of labeled data [8]. This can be done through transfer learning, which uses

DOI: 10.54254/2755-2721/140/2025.21281

knowledge learned in other tasks to help machines learn genetic classification. Second, to reduce the sensitivity to outliers, the data can be cleaned before training to remove outliers and outlier data points. In addition, adopting an ensemble learning approach, allowing multiple machines to work together, can improve the overall accuracy and robustness.

In the field of genetic classification, data set issues are mainly related to the quality of pathological section data, the consistency and privacy protection of global databases, and the insufficient size and diversity of data sets. Pathology section data problems include loss or damage of genetic information, underrepresentation of sections, and high variability of gene expression between individuals. To improve data quality, standardized sample processing and storage protocols can be adopted, and advanced bioinformatics tools can be used to correct data bias [9]. At the same time, by implementing multi-region sampling and multi-sample aggregation techniques, sample representativeness can be enhanced to ensure that samples are fully representative of tissue or individual gene expression [10]. Global database issues include data inconsistencies, privacy protection and compliance with ethical standards, and lack of timely database updates. To comply with privacy and ethical standards, personal genetic information should be handled in strict compliance with relevant standards, and data should be protected using encryption and anonymization techniques. At the same time, an automated data update mechanism should be established to encourage researchers to regularly submit their latest research results to public databases to maintain the timeliness of the databases [11]. The problem of data set size and diversity relates to insufficient data set coverage and lack of representativeness. To expand the dataset size, data collection, and annotation can be used to increase the dataset size, and automated or semi-automated methods can be used for data annotation [12]. To improve the diversity of data sets, a data-sharing platform should be established to encourage data sharing, and international cooperation should be carried out to exchange data resources from different countries to obtain more comprehensive data sets [13].

5. Conclusion

This review discusses the application of AI in gene classification, deeply analyzes the application characteristics and performance of deep learning models such as MLP, CNN, and RNN, and pays special attention to the advantages of sci-kit-learn, BERT and DNASimCLR frameworks in gene classification. Through experiments, these methods have shown remarkable effectiveness in improving the accuracy of gene classification. At the same time, the source and processing methods of the data set are discussed, including pathological section analysis and global database information invocation. The paper points out the limitations of current methods, such as training efficiency, outlier sensitivity, and data dependence, and proposes future research directions such as regularization techniques, pre-training models, and increasing data diversity to improve model performance and robustness. This review guides the development of gene classification techniques and helps to promote the progress of related technologies.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] *Su, Y., Xiang, H., Xie, H., Yu, Y., & Zhao, N. (2020). Application of bert to enable gene classification based on clinical evidence. BioMed Research International, 2020, 1-13.*

[2] *Mallick, P. K., Mohapatra, S. K., Chae, G. S., & Mohanty, M. N. (2020). Convergent learning–based model for leukemia classification from gene expression. Personal and Ubiquitous Computing (4), 1-8.*

[3] *Beyreli, I., Karakahya, O., & Cicek, A. E. (2022). DeepND: Deep multitask learning of gene risk for comorbid neurodevelopmental disorders. Patterns, 3(7).*

DOI: 10.54254/2755-2721/140/2025.21281

[4] Zeng, J., Wu, Q. W., Li, D. H., & et al. (2024, December 2). Aging gene feature selection based on machine learning and classification. Experimental Science and Technology, 1-6.

[5] Yang, M., Wang, Z., Yan, Z., Wang, W., Zhu, Q., & Jin, C. (2024). Dnasimclr: a contrastive learning-based deep learning approach for gene sequence data classification. BMC Bioinformatics, 25(1), 1-13.

[6] Jiang, Q., Fu, X., Yan, S., Li, R., & Grima, R. (2021). Neural network aided approximation and parameter inference of non-markovian models of gene expression. Nature Communications, 12(1).

[7] Xu, J., Zhang, Y., & Zhao, H. (2021). Robustness of machine learning generalization capability on small databases. Journal of Xinjiang Normal University: Natural Science Edition.

[8] Hu, J., Li, X., Hu, G., & et al. (2020). Using neural networks for iterative transfer learning for clustering and cell type classification in single-cell RNA-seq analysis. Nature Machine Intelligence, 2, 607-618.

[9] Yang, M., Wang, Z., Yan, Z., & et al. (2024). DNASimCLR: A deep learning method for the classification of gene sequence data based on contrast learning. BMC Bioinformatics, 25, 328.

[10] Yang, C., Zhang, S., Cheng, Z., Liu, Z., Zhang, L., & Jiang, K., et al. (2022). Multi-region sequencing with spatial information enables accurate heterogeneity estimation and risk stratification in liver cancer. Genome Medicine.

[11] Liu, J. (2024). Bioethical and legal issues of CRISPR/Cas9 gene editing technology. Molecular Plant Breeding, 22(10), 3188-3194.

[12] Yang Z P. (2015). Research on classification method of unbalanced data based on neural network. (Doctoral dissertation, East China University of Science and Technology).

[13] Cong, W., & Xiaoxuan, L. (2020). International Practice of Open Research Data and Its Enlightenment to China. Bulletin of Chinese Academy of Sciences (Chinese Version), 35(1), 11-18.