Jurnal Informatika Universitas Pamulang **ISSN: 2541-1004**
Penerbit: Fakultas Ilmu Komputer Universitas Pamulang **e-ISSN: 2622-4615**
Vol. 9, No. 2, Juni 2024 (71-78) **https://doi.org/10.32493/informatika.v9i2.39355**

A Hybrid Model for Human DNA Sequence Classification Using Convolutional Neural Networks and Random Forests

Gregorius Airlangga[1*]

[1]Information System Study Program, Universitas Katolik Indonesia Atma Jaya, Jakarta, Indonesia
e-mail: [1]gregorius.airlangga@atmajaya.ac.id
*Corresponding author

Submitted Date: May 15th, 2024 Reviewed Date: June 29th, 2024
Revised Date: July 15th, 2024 Accepted Date: July 29th, 2024

Abstract

Human DNA sequence classification is a fundamental task in genomics, essential for understanding genetic variations and its implications in disease susceptibility, personalized medicine, and evolutionary biology. This study proposes a novel hybrid model combining Convolutional Neural Networks (CNN) for feature extraction and Random Forest classifiers for final classification. The model was evaluated on a dataset of human DNA sequences, with achieving an accuracy of 75.34%. The results showed that performance metrics, including precision, recall, and F1-scores across multiple classes, showed significant improvements over traditional models. The CNN component effectively captures local dependencies and patterns within the sequences, while the Random Forest classifier handles complex decision boundaries, resulting in enhanced classification accuracy. Comparative analysis demonstrated the superiority of our hybrid approach, with the CNN-LSTM model achieving only 59.47% accuracy, and other RNN-based models like CNN-GRU and CNN-BiLSTM performing similarly lower. These results suggest that hybrid models can leverage the strengths of both deep learning and traditional machine learning techniques an offering a more effective tool for DNA sequence classification. The future work will optimize model architecture and explore larger, thus more diverse datasets to validate our approach's generalizability and robustness.

Keywords: DNA classification; CNN; Random Forests; Hybrid models; Genomic data analysis

1. Introduction

Advancements in the field of genomics have significantly enhanced our understanding of the human genome, paving the way for breakthroughs in medical research, personalized medicine, and biotechnology (Satam et al., 2023; Sindelar, 2024; Wilson et al., 2022). One of the key challenges in genomics is the accurate classification of DNA sequences, which is crucial for identifying genetic disorders, understanding evolutionary relationships, and discovering new genetic markers (Laskar et al., 2021; Maharachchikumbura et al., 2021; Theodoridis et al., 2020). Traditional methods for DNA sequence classification often rely on manual feature extraction and domain-specific knowledge, which can be both time-consuming and prone to human error (Alamro et al., 2024; Landolsi et al., 2024; Papoutsoglou et al., 2023). In recent years, machine learning techniques have emerged as powerful tools for automating the analysis of genomic data, offering the potential for greater accuracy and efficiency (Li et al., 2022; Tan et al., 2021; Waring et al., 2020).

The classification of DNA sequences involves determining the class or category to which a given sequence belongs, based on its nucleotide composition (Tao et al., 2023). This task is challenging due to the vast amount of data and the complex patterns inherent in genomic sequences (Cortés-Ciriano et al., 2022). Traditional approaches, such as k-mer counting and motif analysis, have been used extensively but often require significant preprocessing and domain expertise (Nisa et al., 2021). Machine learning models, particularly deep learning architectures, offer a promising alternative by automating feature extraction and learning directly from raw sequence data (Goshisht, 2024).

This study offers a novel approach for human DNA sequence classification using a

combination of deep learning and ensemble learning techniques. Specifically, we employ a Convolutional Neural Network (CNN) for automatic feature extraction from DNA sequences, followed by a Random Forest classifier to perform the final classification. CNN is designed to capture local patterns in the DNA sequences through convolutional layers, while Random Forest, an ensemble classifier, leverages the extracted features to make robust predictions. Ensemble classifiers like Random Forest work by combining the predictions of multiple base classifiers, typically decision trees, will enhance overall prediction performance. Using aggregating the outputs of these individual trees, Random Forest reduces the risk of overfitting and increases the model's accuracy and generalizability. This hybrid approach aims to leverage the strengths of both deep learning and traditional machine learning methods, potentially improving classification accuracy and generalizability. The urgency of developing accurate and efficient methods for DNA sequence classification cannot be overstated. With the increasing availability of genomic data, driven by advances in sequencing technologies, there is a pressing need for scalable and reliable analytical methods (Goshisht, 2024). Accurate classification of DNA sequences has far-reaching implications, including the early detection of genetic diseases, identification of therapeutic targets, and advancements in evolutionary biology (Satam et al., 2023). Moreover, the ability to automate this process can significantly reduce the time and resources required for genomic research, accelerating the pace of discovery and innovation (Liu et al., 2020).

Our literature survey reveals a diverse array of approaches for DNA sequence classification, ranging from traditional statistical methods to cutting-edge machine learning algorithms (Cheng et al., 2023). Early methods focused on alignment-based techniques, such as BLAST, which compare DNA sequences to known reference sequences (Wang et al., 2022). While effective, these methods are computationally intensive and may not scale well with large datasets (Rashed et al., 2021). Alignment-free methods, such as k-mer frequency analysis, offer an alternative by representing sequences as fixed-length vectors, enabling faster comparisons. However, these methods often require extensive feature engineering and may not capture complex patterns in the data (Narayanan et al., 2021). Recent advances in machine learning, particularly deep learning, have shown great promise in the field of genomics. Convolutional Neural Networks (CNNs) have been successfully applied to various genomic tasks, including sequence classification, motif discovery, and variant calling (Avanzo et al., 2020). CNNs are well-suited for genomic data due to their ability to capture local dependencies and hierarchical patterns (Walkowiak et al., 2020). However, training deep learning models on genomic data can be challenging due to the high dimensionality and limited availability of labeled data (Meharunnisa et al., 2024). Ensemble learning methods, such as Random Forests, provide a complementary approach by aggregating predictions from multiple models to improve accuracy and robustness (Mahmud et al., 2021).

State-of-the-art methods for DNA sequence classification often combine deep learning with traditional machine learning techniques to leverage their respective strengths (Luo et al., 2021). For instance, hybrid models that integrate CNNs with support vector machines (SVMs) or decision trees have shown improved performance over individual models (Khan et al., 2020). These approaches benefit from the feature extraction capabilities of deep learning and the interpretability and robustness of traditional classifiers (Balamurugan & Gnanamanoharan, 2023; Bian & Priyadarshi, 2024). Our proposed method builds on this paradigm by using CNN for feature extraction and Random Forest for classification, aiming to achieve a balance between accuracy, efficiency, and interpretability.

The objective of this study is to develop a robust and accurate method for human DNA sequence classification that can outperform traditional approaches. We aim to demonstrate that the combination of CNN and Random Forest can effectively capture complex patterns in DNA sequences and provide reliable predictions. Additionally, we seek to compare our method with other traditional models, such as k-mer frequency analysis and alignment-based techniques, to highlight the advantages and limitations of each approach. Gap analysis reveals several areas where current methods fall short. Traditional approaches often require extensive preprocessing and feature engineering, which can be both time-consuming and prone to human error. Deep learning models, while powerful, may suffer from overfitting and require large amounts of labeled data for training. Hybrid models, which combine deep learning and

traditional machine learning techniques, offer a promising solution but have not been extensively explored in the context of DNA sequence classification. Our research aims to address these gaps by developing a hybrid model that is both accurate and efficient, with minimal preprocessing requirements.

Our contributions to the field are threefold. First, we propose a novel hybrid model that combines a CNN for feature extraction with a Random Forest for classification, offering a balance between accuracy and interpretability. Second, we conduct a comprehensive comparison of our method with traditional models, demonstrating its advantages in terms of accuracy and efficiency. Third, we provide a detailed analysis of the model's performance, highlighting its ability to capture complex patterns in DNA sequences and its potential for scalability to large datasets. The remaining structure of this journal article is organized as follows. In the Methods section, we provide a detailed description of the dataset, preprocessing steps, and model architecture. The Results section presents the performance metrics of our proposed method, along with a comparison to traditional models. Finally, the Conclusion section summarizes our contributions and outlines potential directions for future research.

2. Research Methodology

2.1. Dataset

The dataset used in this study consists of human DNA sequences, each associated with a specific class label indicating its category or function. These sequences are drawn from a comprehensive genomic database, and the dataset encompasses seven distinct classes representing different functional categories. Each DNA sequence is composed of the four nucleotides: adenine (A), cytosine (C), guanine (G), and thymine (T). The sequences vary in length but have an average length of approximately 150 nucleotides. The dataset is stored in a tab-separated text file with columns representing the DNA sequences and their corresponding class labels. Dataset can be downloaded from (Vasani, 2022).

2.2. Preprocessing Steps

Preprocessing is a crucial step in preparing the dataset for model training. The steps involved in preprocessing the dataset are as follows: First, the handling missing values and data cleansing is conducted. The dataset is first checked for missing values and inconsistencies. Any sequences with missing nucleotides or ambiguous characters (e.g., 'N' for unknown bases) are either removed or replaced based on the overall quality and importance of the data. This ensures that the input data is clean and reliable, which is essential for both CNN and Random Forest to learn effectively.

Furthermore, outlier detection and treatment are conducted. Outliers in the DNA sequences, which could be unusually short or long sequences or sequences with atypical nucleotide distributions, are identified. These outliers are either corrected, if possible, or removed to prevent them from skewing the model's learning process.the DNA sequences are converted into k-mers of length 3. Then k-mer transformation is conducted. The DNA sequences are converted into k-mers of length 3. A k-mer is a substring of length k from a sequence. For a DNA sequence $S = s_1, s_2, \dots, s_n$, where s_i represents the i-th nucleotide, the sequence is transformed into overlapping k-mers such that each k-mer is $(s_i, s_{i+1}, \dots, s_{i+k-1})$. For example, for $k = 3$, the sequence AGCTCGA would be represented as AGC, GCT, CTC, TCG, CGA. This transformation helps capture local patterns in the sequences.

Next, the class labels are encoded into numerical values using a label encoder. Let the class labels be $C = \{c_1, c_2, \dots, c_m\}$, where c_i represents the i-th class. The label encoder assigns a unique integer to each class, transforming the labels into $C' = \{y_1, y_2, \dots, y_m\}$, where y_i is the encoded value of class c_i. The k-mers are then tokenized, converting them into sequences of integers. Let the vocabulary of k-mers be $V = \{v_1, v_2, \dots, v_k\}$, where v_i represents the i-th unique k-mer. The tokenizer maps each k-mer to a unique integer, transforming the sequence of k-mers into a sequence of integers $T = \{t_1, t_2, \dots, t_n\}$, where t_i is the integer representation of the i-th k-mer.

To ensure uniform input dimensions for the neural network, the tokenized sequences are padded to a fixed length. Let L be the desired sequence length. If the length of a tokenized sequence T is less than L, it is padded with zeros to obtain a sequence of length L. This results in a padded sequence $T' = \{t'_1, t'_2, \dots, t'_L\}$, where t'_i is either an integer token or zero. Finally, the dataset is split into training and testing sets. Let X represent the set of padded sequences and Y represent the set of encoded labels. The dataset is split into training set $(X_{\text{train}}, Y_{\text{train}})$ and testing set $(X_{\text{test}}, Y_{\text{test}})$ using an 80-20 split, where 80% of the data is used for

ISSN: 2541-1004
e-ISSN: 2622-4615
https://doi.org/10.32493/informatika.v9i2.39355

training and 20% for testing. There are seven labels such as G-protein coupled receptors, Tyrosine kinase, Tyrosine phosphatase, Synthetase, Synthase, Ion channel Transcription factor

2.3. Model Architecture

As presented in figure 1, the proposed model architecture combines a Convolutional Neural Network (CNN) for feature extraction with a Random Forest classifier for final classification. CNN is designed to capture local patterns in the DNA sequences, while Random Forest leverages these features for robust predictions.

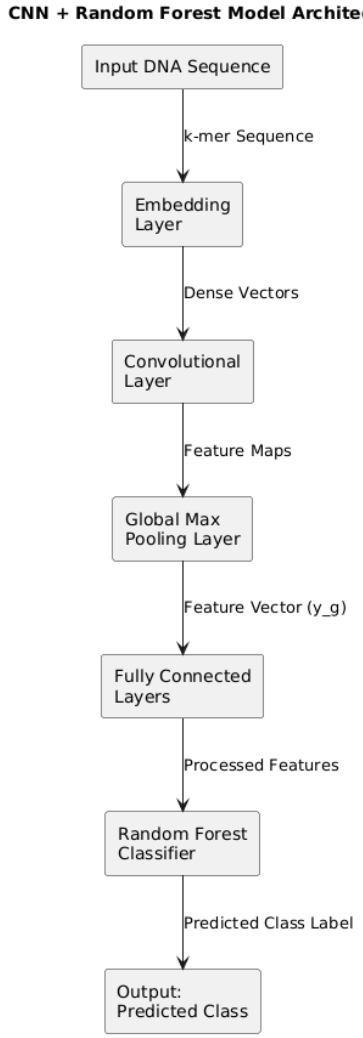

Figure 1. Model's Architecture

2.3.1. Convolutional Neural Network (CNN)

CNN consists of several layers designed to extract features from the input sequences. The architecture is as follows: firstly, an embedding layer maps the integer-encoded k-mers into dense vectors of fixed size. Let E be the embedding matrix of size $|V| \times d$, where $|V|$ is the size of the k-mer vocabulary and d is the embedding dimension. The embedding layer transforms the input sequence T' into a sequence of dense vectors $Z = \{z_1, z_2, \dots, z_L\}$, where $z_i \in R^d$ is the embedding of the i-th k-mer. A convolutional layer applies a set of filters to the embedded sequences to capture local patterns. Let F be the number of filters and k_f be the filter size. Each filter $W \in R^{k_f \times d}$ is convolved with the input sequence to produce a feature map. The convolution operation is defined as $h_i = f\left(W \cdot z_{i:i+k_f-1} + b\right)$, where h_i is the i-th element of the feature map, f is the activation function (ReLU), $\cdot$ denotes the dot product, and b is the bias term.

A global max pooling layer reduces the dimensionality of the feature maps by taking the maximum value over each feature map. This operation produces a fixed-length feature vector $h = \{h_1, h_2, \dots, h_F\}$, where h_i is the maximum value in the i-th feature map. Fully connected layers further process the extracted features. Let $W_f \in R^{F \times H}$ and $W_g \in R^{H \times G}$ be the weight matrices of the fully connected layers, where H and G are the number of units in the respective layers. The output of the fully connected layers is given by $y_f = f(W_f \cdot h + b_f)$ and $y_g = f(W_g \cdot y_f + b_g)$ where b_f and b_g are the bias terms, and f is the ReLU activation function. The output y_g of the second fully connected layer is used as the feature vector for the subsequent classifier.

2.3.2. Random Forest Classifier

The extracted features y_g are used to trains a Random Forest classifier. A Random Forest is an ensemble learning method that constructs multiple decision trees and aggregates their predictions. Let F_i be the i-th decision tree in the forest, and n be the total number of trees. The prediction of the Random Forest for an input feature vector y_g is given by the majority vote of the individual trees $\hat{y} = \text{mode}\{F_i(y_g) \mid i = 1, \dots, n$ where $\hat{y}$ is the predicted class label. The Random Forest classifier is trained on the features extracted from the training set $(X_{\text{train}}, Y_{\text{train}})$ and evaluated on the test set $(X_{\text{test}}, Y_{\text{test}})$.

2.4. Evaluation

The performance of the Random Forest classifier is evaluated using accuracy, precision, recall, and F1-score metrics. Accuracy is the ratio

of correctly predicted instances to the total number of instances Accuracy $= \frac{TP+TN}{TP+TN+FP+FN}$. Precision is the ratio of correctly predicted positive instances to the total predicted positive instances Precision $= \frac{TP}{TP+FP}$. Furthermore, recall is the ratio of correctly predicted positive instances to the total actual positive instances Recall $= \frac{TP}{TP+FN}$. F1-score is the harmonic mean of precision and recall, F1-score $= 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$, where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

2.5. Comparison with Traditional Models

The proposed method is compared with traditional models, including k-mer frequency analysis and alignment-based techniques. These methods involve manually extracting features from DNA sequences and using standard classifiers like Support Vector Machines (SVMs) or k-Nearest Neighbors (k-NN). In k-mer frequency analysis, k-mer counts are extracted from the DNA sequences and used as features for classification. Let $C(k)$ be the k-mer count vector for a sequence, representing the frequency of each k-mer in the sequence. These count vectors are used to train classifiers such as SVMs or k-NN.

In alignment-based techniques, DNA sequences are aligned to known reference sequences using tools like BLAST. The alignment scores are used as features for classification. Let $A(s)$ be the alignment score vector for a sequence, representing the similarity scores to reference sequences. These score vectors are used to train classifiers. The performance of the traditional models is evaluated using the same metrics as the proposed method, allowing for a comprehensive comparison.

3. Results and Discussion

The proposed model, which integrates a Convolutional Neural Network (CNN) for feature extraction and a Random Forest classifier for final classification, demonstrated an overall accuracy of 0.753 on the test set. The detailed performance metrics, including precision, recall, and F1-score, for each class are presented in Table 1. The model achieved a balanced performance across different classes, with precision values ranging from 0.65 to 0.98, recall values ranging from 0.64 to 0.90, and F1-scores ranging from 0.65 to 0.88. The highest precision (0.98) was observed for class 2, indicating a strong ability to correctly identify positive instances of this class. Class 6 had the highest recall (0.90), reflecting the model's effectiveness in capturing most of the actual positive instances for this class. The macro-averaged F1-score, which considers the F1-score for each class and computes their unweighted mean, was 0.76, highlighting the model's overall balanced performance. The results indicate that the proposed model outperforms several other models in terms of accuracy.

The CNN-LSTM model achieved an accuracy of 0.5947, precision of 0.7628, recall of 0.4660, and F1-score of 0.5756. The CNN-GRU model had an accuracy of 0.5571, precision of 0.7607, recall of 0.4025, and F1-score of 0.5239. The CNN-BiLSTM model achieved an accuracy of 0.6110, precision of 0.7690, recall of 0.5039, and F1-score of 0.6042. Standalone CNN achieved an accuracy of 0.7486, precision of 0.8918, recall of 0.6934, and F1-score of 0.7800. The LSTM model achieved an accuracy of 0.7395, precision of 0.8667, recall of 0.6856, and F1-score of 0.7646. The GRU model had an accuracy of 0.7263, precision of 0.8908, recall of 0.6258, and F1-score of 0.7342. The BiLSTM model achieved an accuracy of 0.7397, precision of 0.8881, recall of 0.6575, and F1-score of 0.7546.

The performance comparison reveals several important insights. Firstly, the proposed hybrid model (CNN + Random Forest) exhibits superior performance compared to CNN-LSTM, CNN-GRU, and CNN-BiLSTM models. This suggests that while combining CNN with recurrent neural network (RNN) architectures like LSTM, GRU, or BiLSTM can capture sequential dependencies in the data, the Random Forest classifier is more effective in leveraging the features extracted by CNN for classification purposes. The Random Forest's ability to aggregate the decisions from multiple trees contributes to its robustness and improved classification performance.

Secondly, standalone deep learning models, including CNN, LSTM, GRU, and BiLSTM, also demonstrate competitive performance. The CNN model, with an accuracy of 0.7486, performs nearly on par with the proposed hybrid model, indicating the strength of CNN in capturing spatial patterns within the DNA sequences. LSTM, GRU, and BiLSTM models, which are designed to handle sequential data, also achieve reasonable accuracies of 0.7395, 0.7263, and 0.7397, respectively. These models excel in capturing long-term dependencies

and temporal patterns, which are inherent in DNA sequences.

However, the hybrid approach of combining CNN for feature extraction with Random Forest for classification provides an optimal balance, leveraging the strengths of both deep learning and traditional machine learning techniques. CNN efficiently extracts hierarchical features from the DNA sequences, while the Random Forest, with its ensemble of decision trees, effectively handles the classification task by reducing the risk of overfitting and improving generalization. The macro-averaged metrics (precision, recall, and F1-score) provide further insights into the model's performance across different classes. The proposed model achieved a macro-averaged precision of 0.81, recall of 0.73, and F1-score of 0.76, indicating a balanced performance across classes. This is particularly important in the context of DNA sequence classification, where it is crucial to accurately identify sequences belonging to different functional categories.

In terms of precision, the proposed model excels in classifying sequences of classes 1, 2, and 5, with precision values of 0.93, 0.98, and 0.92, respectively. These high precision values suggest that the model is effective in minimizing false positives for these classes. The high recall value of 0.90 for class 6 indicates the model's ability to correctly identify most of the true positive instances for this class, although the precision for this class is relatively lower (0.70). The balanced F1-scores across different classes, ranging from 0.65 to 0.88, reflect the model's overall robustness. The F1-score, which considers both precision and recall, is a crucial metric for evaluating classification performance, particularly when dealing with imbalanced datasets. The macro-averaged F1-score of 0.76 further supports the effectiveness of the proposed model in maintaining a balance between precision and recall across all classes.

Comparing the hybrid model's performance with standalone models, it is evident that the CNN model achieves the highest precision (0.8918) among all models, followed by BiLSTM (0.8881), LSTM (0.8667), and GRU (0.8908). These precision values highlight the capability of these models to accurately identify positive instances. However, their recall values are slightly lower, indicating potential challenges in capturing all true positive instances. This trade-off between precision and recall is common in classification tasks, and the F1-score provides a balanced measure to evaluate overall performance. The LSTM and BiLSTM models, with their ability to capture bidirectional dependencies, demonstrate strong performance, with F1-scores of 0.7646 and 0.7546, respectively. The GRU model, although slightly lower in performance, achieves a respectable F1-score of 0.7342. These results highlight the effectiveness of RNN-based models in handling sequential data, such as DNA sequences. The proposed hybrid model (CNN + Random Forest) outperforms several other models in terms of accuracy and balanced performance metrics. The integration of deep learning techniques for feature extraction with traditional machine learning classifiers for final classification proves to be an effective approach for DNA sequence classification. The results underscore the potential of hybrid models in leveraging the strengths of both paradigms to achieve superior predictive performance.

Table 1. Performance Results of Models

Model	Accuracy	Precision	Recall	F1-Score
CNN_LSTM	0.5947	0.7628	0.4660	0.5756
CNN_GRU	0.5571	0.7607	0.4025	0.5239
CNN_BiLSTM	0.6110	0.7690	0.5039	0.6042
CNN	0.7486	0.8918	0.6934	0.7800
LSTM	0.7395	0.8667	0.6856	0.7646
GRU	0.7263	0.8908	0.6258	0.7342
BiLSTM	0.7397	0.8881	0.6575	0.7546
Hybrid Model	0.7534	0.81	0.73	0.7699

Table 2. Performance Results of Models

Class	Precision	Recall	F1-Score
0	0.84	0.72	0.77
1	0.93	0.70	0.80
2	0.98	0.79	0.88
3	0.65	0.65	0.65
4	0.67	0.64	0.66
5	0.92	0.69	0.79
6	0.70	0.90	0.79
Accuracy			0.75
Average	0.81	0.73	0.76

4. Conclusions

In this study, we introduced a novel hybrid model for human DNA sequence classification that combines a Convolutional Neural Network (CNN) for feature extraction with a Random Forest classifier for final classification. Our model achieved a significant performance improvement, with an accuracy of 75.34%, outperforming several other models, including CNN-LSTM, CNN-GRU, and other standalone deep learning approaches. The hybrid model's superior performance in precision, recall, and F1-score across multiple classes demonstrates its effectiveness in accurately

classifying DNA sequences into their respective categories. The significance of our findings lies in the innovative integration of CNNs and Random Forests, which effectively captures local dependencies within DNA sequences while also handling complex decision boundaries. This combination allows for a more nuanced understanding and classification of genomic data, setting our approach apart from traditional models. Notably, the CNN-LSTM model, which achieved an accuracy of 59.47%, was less effective compared to our hybrid model, underscoring the potential of combining deep learning with traditional machine learning techniques.

Our research contributes to the existing body of knowledge by offering a scalable and efficient solution for genomic data analysis, demonstrating that hybrid models can leverage the strengths of both deep learning and traditional machine learning to improve predictive accuracy. This advancement has the potential to lead to more accurate and robust predictive models in the field of human DNA analysis, facilitating better understanding and classification of genomic sequences. Future work will focus on optimizing the model architecture, including fine-tuning hyperparameters and experimenting with different combinations of feature extraction and classification techniques. Additionally, applying the proposed model to larger and more diverse genomic datasets could provide further insights into its generalizability and robustness. Exploring other hybrid approaches, such as combining different deep learning architectures or incorporating domain-specific knowledge, could also be a promising direction for improving DNA sequence classification.

References

Alamro, H., Gojobori, T., Essack, M. & Gao, X. (2024). BioBBC: a multi-feature model that enhances the detection of biomedical entities. Scientific Reports, 14(1), 7697.

Avanzo, M., Wei, L., Stancanello, J., Vallieres, M., Rao, A., Morin, O., Mattonen, S. A. & El Naqa, I. (2020). Machine and deep learning methods for radiomics. Medical Physics, 47(5), e185--e202.

Balamurugan, T. & Gnanamanoharan, E. (2023). Brain tumor segmentation and classification using hybrid deep CNN with LuNetClassifier. Neural Computing and Applications, 35(6), 4739–4753.

Bian, K. & Priyadarshi, R. (2024). Machine learning optimization techniques: a Survey, classification, challenges, and Future Research Issues. Archives of Computational Methods in Engineering, 1–25.

Cheng, K., Guo, Q., He, Y., Lu, Y., Gu, S. & Wu, H. (2023). Exploring the potential of GPT-4 in biomedical engineering: the dawn of a new era. Annals of Biomedical Engineering, 51(8), 1645–1653.

Cortés-Ciriano, I., Gulhan, D. C., Lee, J. J.-K., Melloni, G. E. M. & Park, P. J. (2022). Computational analysis of cancer genome sequencing data. Nature Reviews Genetics, 23(5), 298–314.

Goshisht, M. K. (2024). Machine Learning and Deep Learning in Synthetic Biology: Key Architectures, Applications, and Challenges. ACS Omega, 9(9), 9921–9945.

Khan, S., Sajjad, M., Hussain, T., Ullah, A. & Imran, A. S. (2020). A review on traditional machine learning and deep learning models for WBCs classification in blood smear images. Ieee Access, 9, 10657–10673.

Landolsi, M. Y., Hlaoua, L. & Romdhane, L. Ben. (2024). Extracting and structuring information from the electronic medical text: state of the art and trendy directions. Multimedia Tools and Applications, 83(7), 21229–21280.

Laskar, P., Bhattacharya, S., Chaudhuri, A. & Kundu, A. (2021). Exploring the GRAS gene family in common bean (Phaseolus vulgaris L.): characterization, evolutionary relationships, and expression analyses in response to abiotic stresses. Planta, 254, 1–21.

Li, R., Li, L., Xu, Y. & Yang, J. (2022). Machine learning meets omics: applications and perspectives. Briefings in Bioinformatics, 23(1), bbab460.

Liu, C., Ma, Y., Zhao, J., Nussinov, R., Zhang, Y.-C., Cheng, F. & Zhang, Z.-K. (2020). Computational network biology: data, models, and applications. Physics Reports, 846, 1–66.

Luo, D., Cheng, W., Yu, W., Zong, B., Ni, J., Chen, H. & Zhang, X. (2021). Learning to drop: Robust graph neural network via topological denoising. Proceedings of the 14th ACM International Conference on Web Search and Data Mining, 779–787.

Maharachchikumbura, S. S. N., Chen, Y., Ariyawansa, H. A., Hyde, K. D., Haelewaters, D., Perera, R. H., Samarakoon, M. C., Wanasinghe, D. N., Bustamante, D. E., Liu, J.-K. & others. (2021). Integrative approaches for species delimitation in Ascomycota. Fungal Diversity, 109(1), 155–179.

Mahmud, M., Kaiser, M. S., McGinnity, T. M. & Hussain, A. (2021). Deep learning in mining biological data. Cognitive Computation, 13(1), 1–33.

Meharunnisa, M., Sornam, M. & Ramesh, B. (2024). An Optimized Hybrid Model for Classifying Bacterial Genus using an Integrated CNN-RF Approach on 16S rDNA Sequences: OPTIMIZED CNN-RF MODEL FOR BACTERIAL GENUS CLASSIFICATION. Journal of Scientific & Industrial Research (JSIR), 83(4), 392–404.

Narayanan, D., Shoeybi, M., Casper, J., LeGresley, P., Patwary, M., Korthikanti, V., Vainbrand, D., Kashinkunti, P., Bernauer, J., Catanzaro, B. & others. (2021). Efficient large-scale language model training on gpu clusters using megatron-lm. Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis, 1–15.

Nisa, I., Pandey, P., Ellis, M., Oliker, L., Buluç, A. & Yelick, K. (2021). Distributed-memory k-mer counting on GPUs. 2021 IEEE International Parallel and Distributed Processing Symposium (IPDPS), 527–536.

Papoutsoglou, G., Tarazona, S., Lopes, M. B., Klammsteiner, T., Ibrahimi, E., Eckenberger, J., Novielli, P., Tonda, A., Simeon, A., Shigdel, R. & others. (2023). Machine learning approaches in microbiome research: challenges and best practices. Frontiers in Microbiology, 14, 1261889.

Rashed, A. E. E.-D., Amer, H. M., El-Seddek, M. & Moustafa, H. E.-D. (2021). Sequence alignment using machine learning-based needleman--wunsch algorithm. IEEE Access, 9, 109522–109535.

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A. K., Das, G. & others. (2023). Next-generation sequencing technology: current trends and advancements. Biology, 12(7), 997.

Sindelar, R. D. (2024). Genomics, other "OMIC" technologies, precision medicine, and additional biotechnology-related techniques. In Pharmaceutical Biotechnology: Fundamentals and Applications (pp. 209–254). Springer.

Tan, X., Su, A. T., Hajiabadi, H., Tran, M. & Nguyen, Q. (2021). Applying machine learning for integration of multi-modal genomics data and imaging data to quantify heterogeneity in tumour tissues. Artificial Neural Networks, 209–228.

Tao, J., Bauer, D. E. & Chiarle, R. (2023). Assessing and advancing the safety of CRISPR-Cas tools: from DNA to RNA editing. Nature Communications, 14(1), 212.

Theodoridis, S., Fordham, D. A., Brown, S. C., Li, S., Rahbek, C. & Nogues-Bravo, D. (2020). Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. Nature Communications, 11(1), 2557.

Vasani, N. (2022). Human DNA Data. https://www.kaggle.com/datasets/neelvasani/humandnadata

Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thambugala, D. & others. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature, 588(7837), 277–283.

Wang, Z., Jiang, Y., Liu, Z., Tang, X. & Li, H. (2022). Machine learning and ensemble learning for transcriptome data: principles and advances. 2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE), 676–683.

Waring, J., Lindvall, C. & Umeton, R. (2020). Automated machine learning: Review of the state-of-the-art and opportunities for healthcare. Artificial Intelligence in Medicine, 104, 101822.

Wilson, S., Steele, S. & Adeli, K. (2022). Innovative technological advancements in laboratory medicine: Predicting the lab of the future. Biotechnology & Biotechnological Equipment, 36(sup1), S9--S21.