RESEARCH ARTICLE

DeepECD: A model for predicting plant extrachromosomal circular DNA from sequences

Jiming Hu, Zhongzhen Tang, Yongzhen Wang, Jun Yan, Xiaoyong Sun*

College of Information Science and Engineering, Shandong Agricultural University, Taian, Shandong, China.

Received: January 9, 2024; **accepted:** February 29, 2024.

Extrachromosomal circular DNA (eccDNA) is a double-stranded circular DNA originating from chromosomes, constituting a vital element of the genetic system. Many researchers have found that eccDNA plays an important role in plant growth and development. Artificial intelligence methods are now widely used in the field of bioinformatics for various gene sequences. However, there is no deep learning model to predict plant eccDNA with high accuracy. This study developed DeepECD deep learning model in Python language and compared DeepECD with commonly used machine learning and deep learning models to evaluate its performance in predicting plant eccDNA. The results showed that the gene sequence length of 500 bp with the first 250 bp from the upstream boundary of the eccDNA gene coding region and the last 250 bp from the downstream boundary of the eccDNA gene coding region trained using One-hot encoding method had the highest accuracy in DeepECD. Models trained by different tissues from the same species were able to generalize. However, models trained by different species did not generalize. The results confirmed that plant eccDNA could be accurately predicted by DeepECD. The result of this study could speed up research related to plant eccDNA identification and fill the gap for the application of artificial intelligence methods within the field of plant eccDNA. The resulting model provided a new predictive tool for the study of plant eccDNA and expedited the research process related to plant eccDNA identification.

Keywords: extrachromosomal circular DNA; plant eccDNA; deep learning; machine learning.

***Corresponding author:** Xiaoyong Sun, College of Information Science and Engineering, Shandong Agricultural University, Taian 271018, Shandong, China. Email: sunx1@sdau.edu.cn.

Introduction

Extrachromosomal circular DNA (eccDNA) is a double-stranded circular DNA originating from chromosomes, constituting a vital element of the genetic system [1], and is widely found in different eukaryotes [2, 3]. However, due to its unique structure, its function remains a mystery. With the continuous development of high-throughput technology, numerous research papers published since 2012 have confirmed that eccDNA exists in large quantities and has important biological functions. It not only serves specific roles in development, aging, and evolution, but also assumes a crucial regulatory role in disease expression [4]. The loss of eccDNA homeostasis facilitates tumor initiation, malignant progression, and heterogeneous evolution in many cancers [5, 6]. Hence, eccDNA is an important link in analyzing the gene regulatory network. It is expected to be a potential biomarker for medical diagnosis and may play an important role in disease treatment [4, 7-9]. In plants, high eccDNA load may alter

DNA repair pathways leading to genome instability [10], and it participates in the regulation of male plant sterility and directly affects the process of plant breeding [11]. eccDNA in eukaryotes includes DNA from organelles, such as chloroplasts and mitochondria, as well as small polydisperse circular DNAs, amplified genes, and so on. This type of DNA originates mainly from repetitive sequences in the genome, replicates independently of the chromosomes, and is widespread in the genomes of eukaryotes [3]. Researchers have now detected the presence of this DNA in yeast, pigeons, humans, rodents, Xenopus, and plants [2, 3, 12, 13]. eccDNA can be produced at any location in the genome and can consist of hundreds of base pairs (bp) or even a few megabases (Mb) [5]. Based on the available studies, eccDNA is likely a key link in biological evolution and plasticity.

When dealing with large-scale genomic data, traditional methods of gene sequence analysis often face multiple challenges. With the development of information technology, deep learning has achieved predictive performance comparable to humans in several area tasks such as image processing and natural language processing [14, 15]. Given the limitations of traditional laboratory methods and the superiority of artificial intelligence methods, many researchers have developed artificial intelligence methods to analyze gene sequences. Some researchers have achieved excellent results regarding promoter prediction [16], circular RNA identification [17], and protein structure comparison [18] by using artificial intelligence. Abbasi *et al*. developed iLEC-DNA, a machine learning prediction model, to predict long eccDNA sequences [19]. Chang *et al*. developed DeepCircle, a deep learning model, to predict short eccDNA in humans [20]. However, there has not yet been a deep learning model developed to predict plant eccDNA. In addition, the existing models are only for long or short eccDNA. There is no deep learning model that predicts both long and short eccDNA with good performance.

This study aimed to establish a high-performance model to predict plant eccDNA. The Python language was applied to develop a deep learning model named DeepECD. The model consisted of multiple convolutional and long short-term memory (LSTM) layers. The model was then optimized using pooling, dropout, and gradient descent. The resulting DeepECD model could expedite the process of research related to plant eccDNA identification and fill the gap in the application of artificial intelligence methods in the field of plant eccDNA. This study would also provide a new predictive tool for plant eccDNA research, as well as a basis and guidance for further exploration in related fields.

Materials and methods

Data mining and processing

Selecting the appropriate dataset is crucial for constructing a high-performance deep learning predictive model for eccDNA. Inappropriate data may cause the model to predict incorrectly or lead to a bias toward positive or negative samples. It may also cause the model to fail to learn information useful for eccDNA classification and thus make incorrect decisions about eccDNA classification. The plant eccDNA data was downloaded from the PlantEccDNA database (http://123.56.104.85/PlantEccDNA/), which contained 2.48 GB of eccDNA-related information including 475,199 *Arabidopsis* eccDNA sequences [21] and 11,638 rice eccDNA sequences [22]. The downloaded information covered (1) the species including *Arabidopsis* and rice, (2) tissue including stems, flowers, roots, leaves for *Arabidopsis* and seeds, healing tissues, leaves for rice, (3) genome coordinates including the chromosome number, the start and stop positions of the eccDNA sequence, and positive and negative strands, and (4) reference genome version. The Bedtools 2.25.0 software (https://github.com/arq5x/bedtools2) was used to extract eccDNA sequences from *Arabidopsis* (*Arabidopsis thaliana*) (TAIR10) and rice (*Oryza sativa Japonica*) (IRGSP-1.0) [23]. To ensure the accuracy of the extracted eccDNA sequences,

Integrative Genomics Viewer (IGV) v2.11.9 (https://igv.org/) was applied for manual inspection of the extracted eccDNA sequences [24]. To establish a high quality and reliable positive eccDNA dataset, 10,000 eccDNA sequences were randomly extracted from the tissue of each species. If the total number of sequences was less than 10,000, all eccDNA sequences were extracted. According to previous studies, CD-hit (http://weizhong-lab.ucsd.edu/cd-hit/) was applied to extract eccDNA with the sequence similarity larger than 80% from each tissue for the positive samples [25-27]. The highly similar eccDNA sequences were clustered together to produce high-quality positive samples that represented eccDNA features. This method prevented certain abnormal sequences from interfering with the model. There were two schemes for generating eccDNA-negative samples. The first one selected sequence from the genomic background [28, 29]. The eccDNA sequences were first removed from the genome, and then gene sequences were randomly extracted from the genomic sequences after removing the eccDNA sequences, and the selected gene sequences did not intersect with the eccDNA sequences. This scheme was also sometimes applied by clustering negative and positive samples and removing closely related negative and positive samples. However, this scheme had some drawbacks that (1) it might result in different distributions of negative and positive samples, (2) training the dataset after applying the clustering method to remove the closely related negative and positive samples might lead to overestimation of the model performance, (3) the model performed well on very different positive samples, but it might perform poorly on some similar samples, which could lead to insufficient model generalization. The second scheme was nucleotide reorganization using fasta_ushuffle (https://github.com/agordon/fasta_ushuffle) [29, 30], which eliminated the drawbacks of scheme one. This scheme could disrupt the biological sequence while preserving the k-let counts, and the method resulted in negative samples that were consistent with the distribution of positive samples. According to our encoding method, the k value was set to 1. The fasta_ushuffle was applied to reorganize the eccDNA of positive samples by accounting to obtain suitable eccDNA-negative samples. For each species, 5% of the positive samples and 5% of the negative samples were extracted to form the test set. In the remaining 90% of the dataset, 10% of the sequences were randomly extracted for validation during model training.

Model architecture

The structure of the constructed DeepECD model was illustrated in Figure 1. DeepECD consisted of two feature extractors and a layer of the LSTM network. Feature extractor 1 consisted of Conv1D_1 (filters = 128, kernel_size = 4) and Conv1D_2 (filters = 64, kernel_size = 4). The final output processed by the maximum pooling and discard layers was used as the output of the first feature extractor. Feature extractor 2 consisted of Conv1D_3 (kernel_size = 4) and Conv1D_4 (kernel_size = 16), where the filters were dynamically changed according to the shape of the Conv1D_2 output vector. The feature vector extracted from Conv1D_3 and Conv1D_4 was processed by global maximum pooling and then summed. The output of feature extractor 2 was shaped according to the shape of the output matrix of feature extractor 1. The processed vectors were spliced with the output vectors. The first feature extractor was used as input to the LSTM network. The result was obtained after two fully connected layers. The DeepECD model structure used L2 regularization for all convolutional layers and the value of 0.1 for all dropout layers. The activation function of the sense layer was sigmoidal.

Loss function

To optimize the model, the combination of binary cross-entropy loss (equation 1-1), and hinge loss (equation 1-2) were chosen as the loss functions for the DeepECD model.

$$loss_{binary\ cross-entropy} = -\frac{1}{N}\sum_{i=1}^{N}[y_i\ log(\hat{y}_i) + (1-y_i)\ log(1-\hat{y}_i))] \quad (1\text{-}1)$$

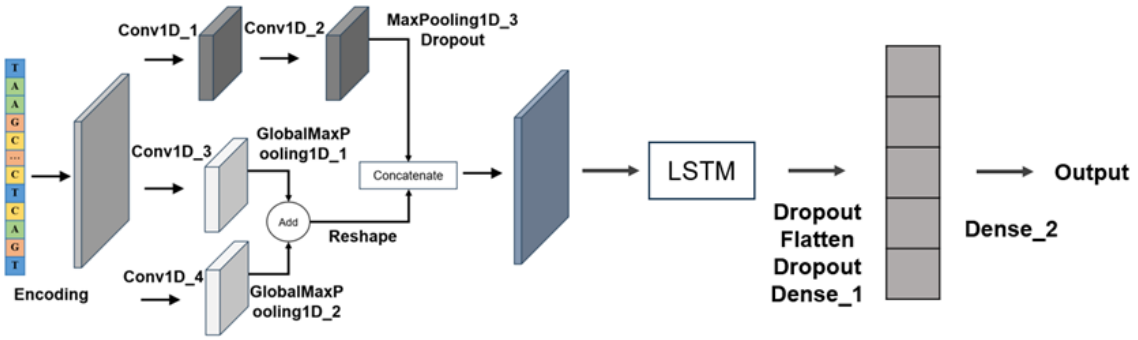


Figure 1. Structure of the DeepECD model.

$$loss_{\text{hinge}} = \frac{1}{N}\sum_{i=1}^{N} max\ (0,1 - \hat{y}_i z_i) \quad (1\text{-}2)$$

The DeepECD loss function formula was shown in equation 1-3 below.

$$loss_{DeepECD} = 0.95 loss_{\text{Binary Crossentropy}} + 0.05\ loss_{\text{Hinge}} \quad (1\text{-}3)$$

where N was the number of samples. $\hat{y}_i$ was the probability that the sequence was predicted to be a positive sample. y_i was the sequence truth category. z_i was the true sample category. The equation differentiated from y_i was y_i = 0 when z_i = -1, and y_i = 1 when z_i = 1.

Evaluation criteria

To evaluate the performance of deep learning models, the commonly used metrics including accuracy (equation 2-1), the F1 score (equation 2-2), and the Matthews correlation coefficient (MCC) (equation 2-3) were employed to evaluate the model performance.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (2\text{-}1)$$

$$F1\ score = \frac{2TP}{2TP+FP+FN} \quad (2\text{-}2)$$

$$MCC = \frac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \quad (2\text{-}3)$$

where TP was the number of positive samples correctly identified. FP was the number of negative samples misreported. TN was the number of negative samples correctly identified. FN was the number of positive samples misreported. Accuracy took a value in the range (0, 1), the F1 score took values in the range (0, 1), and the MCC took values in the range (-1, 1). For all the above metrics, the closer to 1, the better the model performance. When MCC was 1, the model fitted perfectly. However, when it was -1, the model did not fit at all.

Cross-species experimental design

To test if the model could generalize across different species and if enhancing the complexity of species in the training dataset would improve the model's performance, the following experiments were designed, which included (1) training DeepECD with *Arabidopsis* data and tested the model with rice data, (2) training DeepECD with rice data and tested the model with *Arabidopsis* data, (3) randomly extracted 5% of the positive rice sample and 5% of the negative rice sample for testing. A mixture of the remaining 90% of rice data and 10% of randomly extracted *Arabidopsis* data with balanced positive and negative samples was used for training, (4) randomly extracted 5% of the *Arabidopsis* positive sample and 5% of the *Arabidopsis* negative sample for testing. A mixture of the remaining 90% of *Arabidopsis* data and 10% of randomly extracted rice data with balanced positive and negative samples was used for training.

Cross-tissue experimental design

To determine if eccDNA from different tissues shared some common features and if it was possible to recognize eccDNA from different

Table 1. Extrachromosomal circular DNA (eccDNA) data extraction.

Species	Tissue	Number of original sequences	Positive sample (CD hit)	Negative sample (fasta_ushuffle)	All data
Arabidopsis thaliana	Stem	10,000	4,748	4,748	9,496
	Leaf	10,000	4,943	4,943	9,886
	Flower	10,000	5,071	5,071	10,142
	Root	10,000	5,227	5,227	10,454
Rice	Callus	842	358	358	716
	Seed	2,499	2,236	2,236	4,472
	Leaf	8,297	6,560	6,560	13,120

tissues of the same species and from different tissues of different species by using a model trained by eccDNA from one tissue, DeepECD was trained with eccDNA from different tissues and tested with eccDNA from other tissues. Among them, 10% of the eccDNA data used for training was taken to generate the test set. For example, 90% of *Arabidopsis* flower eccDNA was used as the training set and 10% of *Arabidopsis* flower eccDNA and all other tissues were used as the test set, or 90% of *Arabidopsis* leaf tissue eccDNA was used as the training set and 10% of *Arabidopsis* leaf eccDNA and all other tissue eccDNA were used as the test set. 10% of the sequences from the training set were randomly removed and used as the validation set. The test sets with the best model performance were set as the validation set.

Results and discussion

A total of 58,286 sequences were extracted for model training. The sample details for each species and tissue were shown in Table 1.

Model comparison

The DeepECD model was compared with 10 other models including naive bayes [31], decision tree [32], random forest (n_estimators = 100) [33], k-nearest neighbors (n_neighbors = 5) [34], logistic regression (max_iter = 1,000) [35], long short-term memory (LSTM) [36], bidirectional long short-term memory (BiLSTM) [36], gated recurrent unit (GRU) [37], bidirectional gated recurrent unit (BiGRU) [37], and convolutional neural network (CNN) [38]. For eccDNA sequences > 500 bp, the 250bp upstream boundary and the 250bp downstream boundary from the eccDNA coding region were intercepted, respectively, and concatenated them to form DNA sequences with a length of 500 bp. For eccDNA sequences < 500 bp, they were populated from 0 to 500 bp. All eccDNA sequences were encoded using One-hot encoding, and the model that showed the best validation set performance to test the test set during the training process was selected. All deep learning models used for comparison had identical hyperparameters as DeepECD except for the loss function. The binary cross-entropy loss was chosen for comparison of the deep learning models. The deep learning model that showed the best validation set performance for testing was chosen. The accuracy, the F1 score, and the MCC of the models for the test set were then determined (Table 2). The prediction performances of all the machine learning models based on eccDNA were very poor with an accuracy rate of about 50%. In contrast, the deep learning models showed very good performance regarding the recognition of eccDNA with the accuracy and the F1 score all over 0.85 and the MCC larger than 0.7. DeepECD performed better for rice and *Arabidopsis* eccDNA recognition than that of all the other tested deep learning models with both accuracy and the F1 score over 0.91 and the MCC over 0.82 for rice and *Arabidopsis* eccDNA recognition. Because the machine learning models performed poorly, the receiver operating characteristic (ROC) curves for the deep learning models were applied to calculate

Table 2. The results of model comparison between *Arabidopsis thaliana* and rice.

Model	Species	Accuracy	F1 score	MCC
Naive Bayes	Rice	0.5098	0.5244	0.0200
	Arabidopsis	0.5627	0.6214	0.1094
Decision tree	Rice	0.5060	0.4992	0.0120
	Arabidopsis	0.5632	0.5982	0.1196
Random forest	Rice	0.5115	0.4877	0.0226
	Arabidopsis	0.5667	0.5959	0.1291
K-Nearest Neighbor	Rice	0.5098	0.4989	0.0195
	Arabidopsis	0.5427	0.7036	——
Logistic regression	Rice	0.4907	0.4965	-0.0184
	Arabidopsis	0.5554	0.6009	0.1000
LSTM	Rice	0.8695	0.8677	0.7394
	Arabidopsis	0.8847	0.8947	0.7681
BiLSTM	Rice	0.8662	0.8647	0.7327
	Arabidopsis	0.8866	0.8983	0.7701
GRU	Rice	0.8564	0.8540	0.7133
	Arabidopsis	0.8930	0.9022	0.7847
BiGRU	Rice	0.8684	0.8627	0.7395
	Arabidopsis	0.8890	0.8985	0.7767
CNN	Rice	0.8684	0.8636	0.7388
	Arabidopsis	0.8872	0.8975	0.7725
DeepECD	Rice	0.9110	0.9119	0.8222
	Arabidopsis	0.9112	0.9203	0.8201

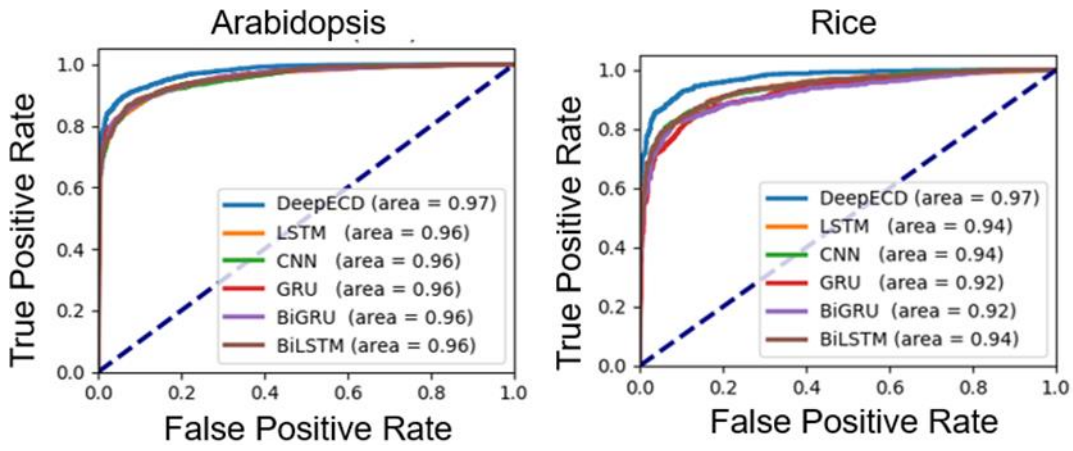
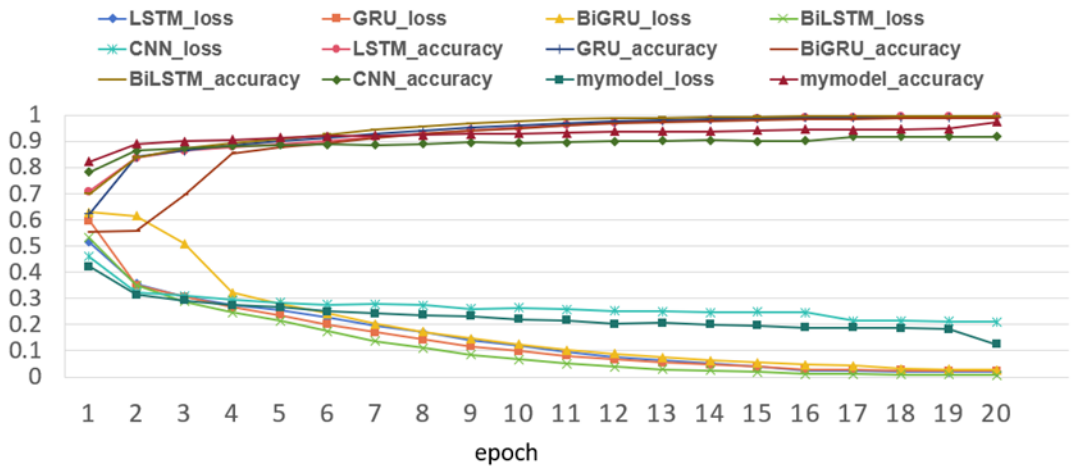


Figure 2. *Arabidopsis thaliana* and rice receiver operating characteristic (ROC) curves and the area under the curve (AUC).

the area under the curve (AUC) (Figure 2). The results showed that DeepECD had an AUC of 0.97 for the recognition of rice and *Arabidopsis* eccDNA, which further verified the superior performance of DeepECD in recognizing eccDNA from these species.

The changes in accuracy, loss, *val* accuracy, and *val* loss per epoch for all deep learning models during training with eccDNA from rice and *Arabidopsis* were then investigated (Figure 3). The accuracy of all deep learning models for the training set gradually approached 1 as the epoch increased. Hence, all deep learning models

A.



B.

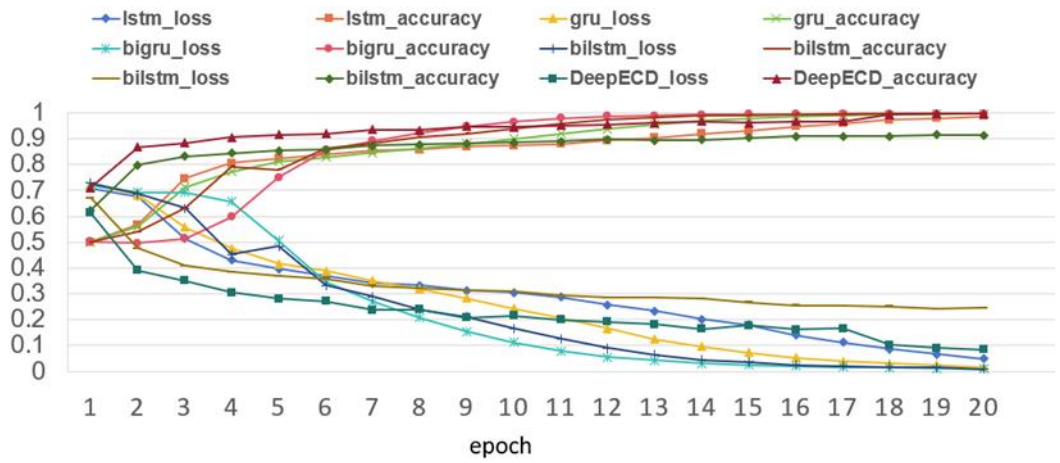


Figure 3. The training results. **A.** *Arabidopsis* training set accuracy and loss for each epoch. **B.** Rice training set accuracy and loss for each epoch.

showed good performance for the recognition of eccDNA based on the training set. The accuracy of DeepECD based on the *Arabidopsis* training set was lower than that of other deep learning models, except for the CNN model that DeepECD had higher accuracy than CNN. Moreover, the loss of DeepECD was greater than that of other models except for the CNN model that DeepECD had lower loss than CNN. For the *Arabidopsis* validation set, DeepECD showed large fluctuations in accuracy and loss at the eighth epoch, which suggested that DeepECD had lower accuracy than that of the other deep learning models but higher accuracy than the BiLSTM model. Also, DeepECD had larger loss than that of the other models and lower loss than that of the BiLSTM model (Figure 4A). For the rice validation set, DeepECD showed larger fluctuations in accuracy and loss at the sixth and sixteenth epochs (Figure 4B). The results suggested that DeepECD did not perform as well as some of the deep learning models in the training set, but it outperformed all deep learning models in the validation set. Hence, DeepECD had better generalization performance.

Analysis of encoding methods

According to recent studies, the encoding method has an impact on the performance of a model [39]. In this study, 10 encoding methods were explored including One-bit encoding, Index encoding, Complementary encoding, BaseN

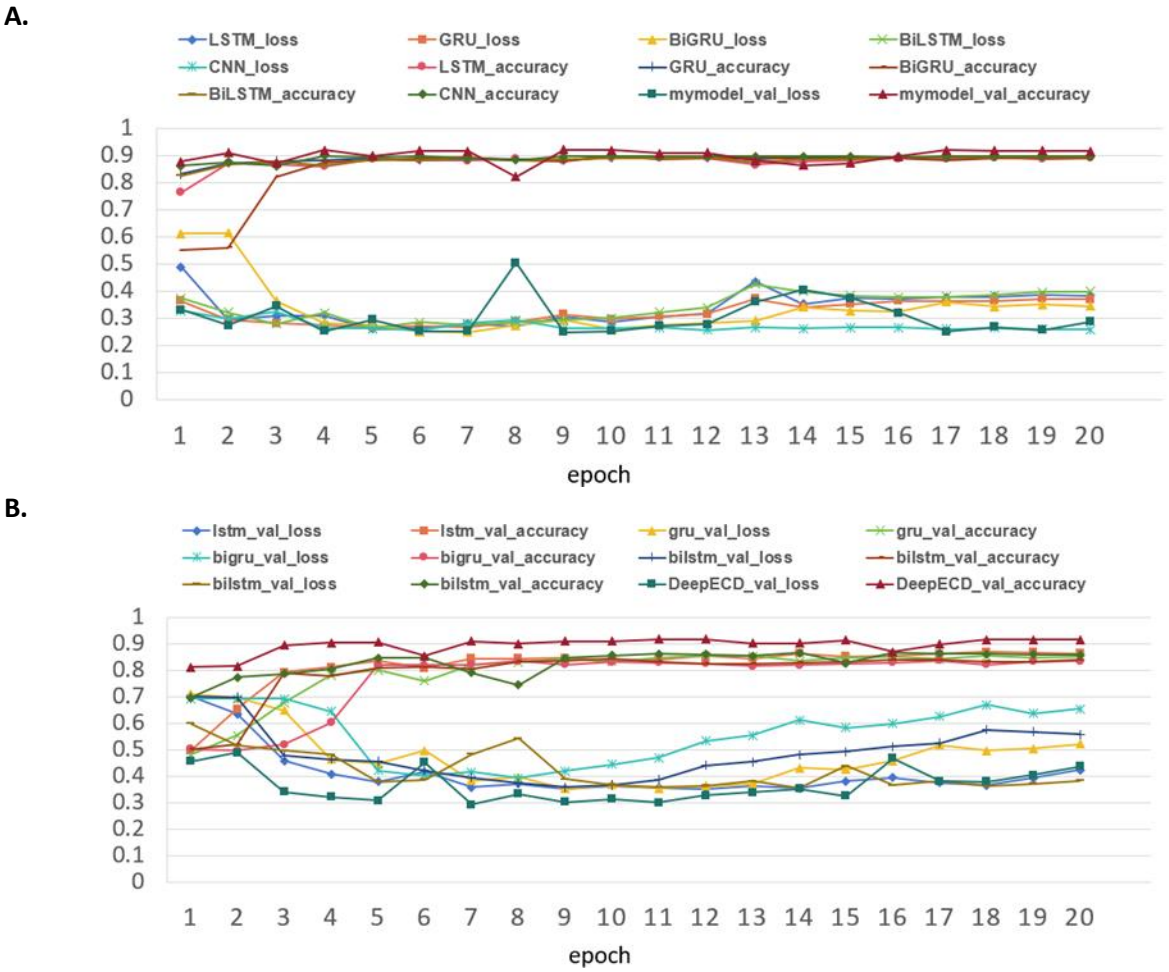


Figure 4. The Validation results. **A.** Plot of the *A. thaliana* validation set accuracy and loss variation with each epoch. **B.** Plot of the rice validation set accuracy and loss variation with each epoch.

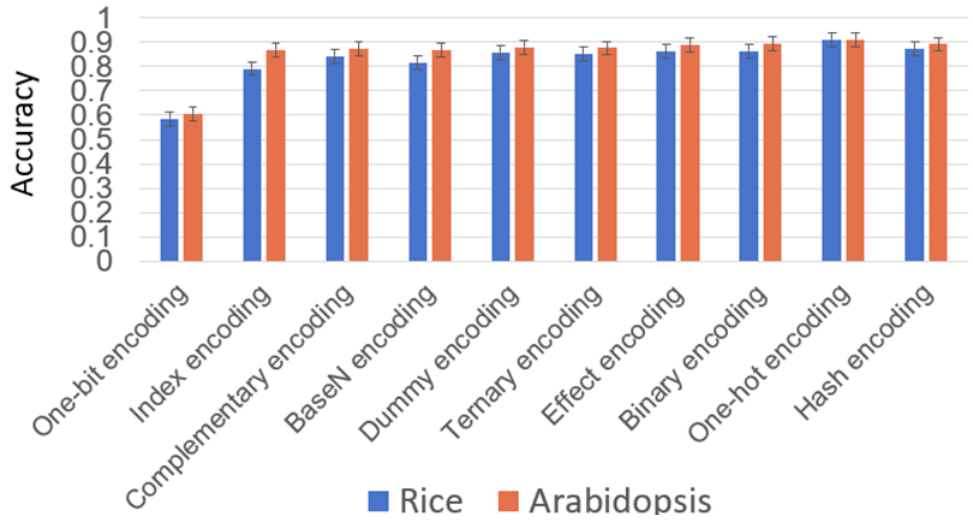


Figure 5. Comparison of the accuracy of different encoding methods of rice and *Arabidopsis* extrachromosomal circular DNA (eccDNA).

Table 3. Species model generalization analysis.

	Training set	Test set	Accuracy	F1 score	Matthew correlation coefficient
Experiment 1	*Arabidopsis*	Rice	0.8208	0.7953	0.6624
Experiment 2	Rice	*Arabidopsis*	0.7239	0.7017	0.4976
Experiment 3	Rice (90%) + *Arabidopsis* (10%)	Rice (10%)	0.9067	0.9042	0.8144
Experiment 4	Rice (10%) + *Arabidopsis* (90%)	*Arabidopsis* (10%)	0.9068	0.9176	0.8109

encoding, Dummy encoding, Ternary encoding, Effect encoding, Binary encoding, One-hot encoding, and Hash encoding for DeepECD. The accuracies of trained DeepECD encoded with each method were shown in Figure 5. One-hot encoding demonstrated the best performance in recognizing eccDNA, followed by Hash encoding and Binary encoding. One-bit encoding showed the worst performance.

Cross-species model performance

The results of cross-species model performance were shown in Table 3. The eccDNA of different species demonstrated some common properties (Experiments 1 and 2) with the model trained by eccDNA from one species could be generalized. The effect of generalization of DeepECD trained by eccDNA of different species was not the same, where DeepECD trained by *Arabidopsis* eccDNA performed better for rice eccDNA than DeepECD trained by rice eccDNA performed for *Arabidopsis* eccDNA. The results of experiments 3 and 4 revealed that the accuracy of mixing *Arabidopsis* and rice data as a training set was 0.8903 to predict rice and 0.9061 to predict *Arabidopsis*. The F1 score and MCC were lower than those of the model trained with a single species. Thus, mixing eccDNA from multiple species did not improve the model's performance in predicting eccDNA from one of the species.

Cross-tissue model performance

When using different tissues from the same species as the training and test sets, the predictive performance of the model was good (Figure 6). The accuracy of the model to predict eccDNA in different tissues from the same species fluctuated around 0.9 for the training and test sets. The accuracy of the model to predict eccDNA in different tissues from a different species fluctuated around 0.8 for the training and test sets. However, the models trained by rice callus as the training set did not perform well, probably because the rice callus dataset was too small for the models to learn the features of eccDNA. Interestingly, the model trained by *Arabidopsis* stem tissue performed better on the other tested tissues than the model trained on itself as the training set. The models trained by *Arabidopsis* leaf and root tissue as the training sets performed better in predicting *Arabidopsis* flower eccDNA than the model trained by *Arabidopsis* flower tissue as the training set. The model trained by rice leaf tissue performed better than the model trained by rice seed in predicting rice seed eccDNA.

Data selection

This study also investigated whether datasets with different intercept lengths affected the model's performance. eccDNA sequences with lengths of 100, 150, 200, 250, 300, 350, 400, 450, and 500 bp at both ends were extracted. For eccDNA with a length greater than 200 bp, 100 bp was intercepted away from the boundary at the front and back ends of the eccDNA sequence and spliced them into a 200 bp sequence. For eccDNAs less than 200 bp in length, the zeros were filled up to 200 bp. For each dataset, 90% of eccDNA sequences were used as the training set and 10% as the test set. From the training set, 10% of the sequences were randomly removed for the validation set. The ratio of positive and negative samples in the training and test sets was 1:1. The model that performed the best was used

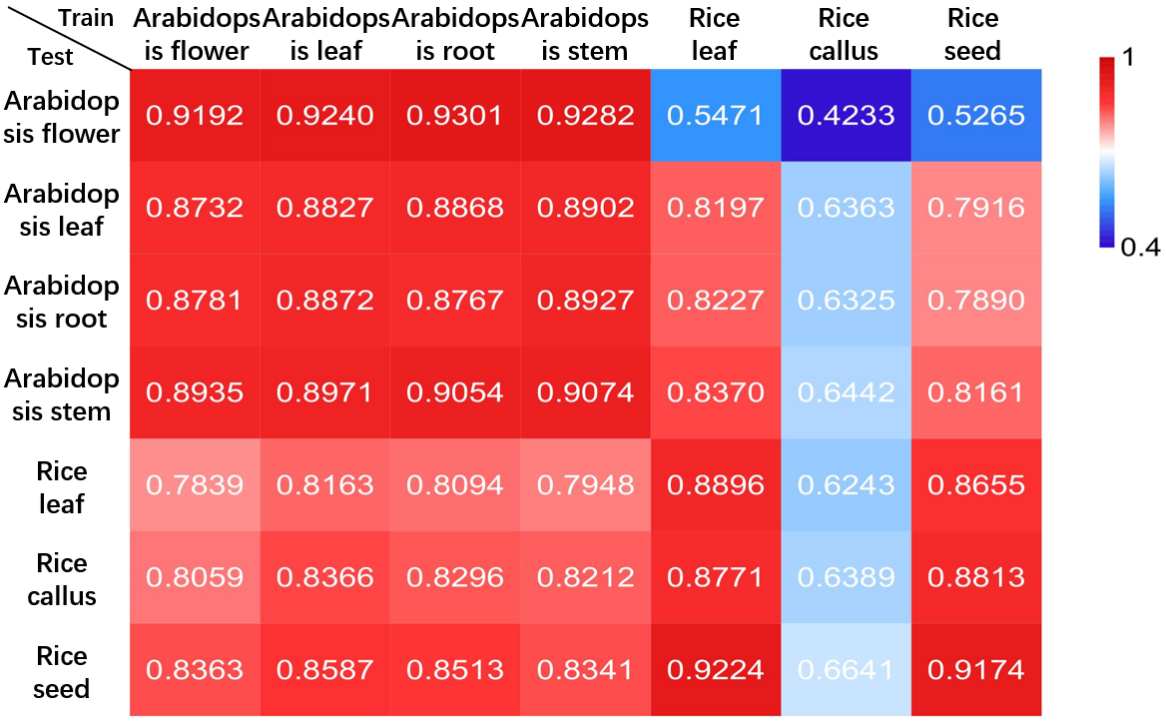


Figure 6. The accuracy of each tissue extrachromosomal circular DNA (eccDNA) model in predicting eccDNA from each tissue.

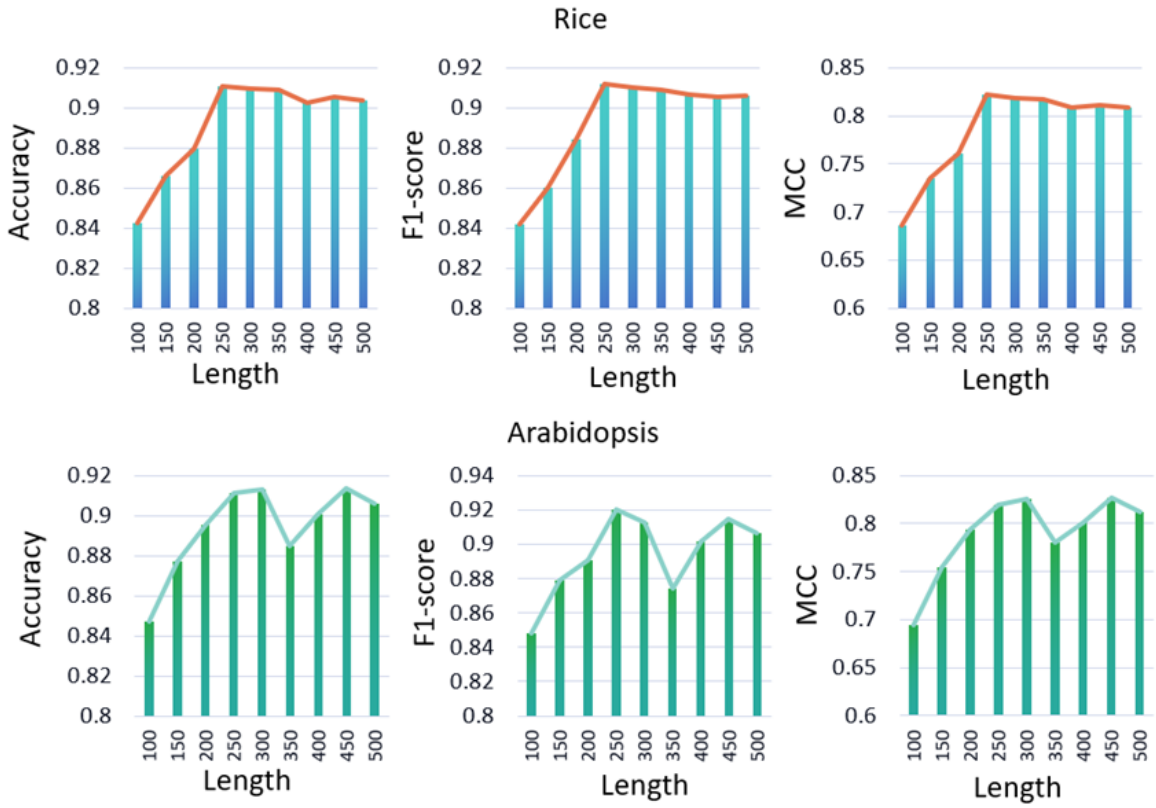


Figure 7. The impact of datasets with different intercept lengths on the model's performance.

for the validation set testing. The results showed that the accuracy, the F1 score, and the MCC gradually increased in rice from the 100 to 250 bp intercept length datasets, while the accuracy, the F1 score, and the MCC gradually smoothed from the 250 to 500 bp intercept length datasets. On the other hand, the accuracy, the F1 score, and the MCC also gradually increased for the 100 to 250 bp intercept length datasets in *Arabidopsis*. However, unlike the rice model, the performance of the *Arabidopsis* model suddenly decreased at the 350 bp intercept length, and then rose again from the 350 to 450 bp intercept length datasets (Figure 7).

Conclusion

This study confirmed that deep learning could predict plant eccDNA. The proposed DeepECD, a plant eccDNA prediction model, consisted of self-constructed feature extractor, loss function, and LSTM network, which showed excellent performance in recognizing rice and *Arabidopsis* eccDNA with accuracy exceeding 0.91. The results showed that DeepECD outperformed not only commonly used deep learning models as well as machine learning models, but also DeepCircle (accuracy = 83.31 ± 4.18%) and iLEC-DNA (accuracy = 67 - 73%). The impacts of different factors including different encoding methods and intercept lengths on the model's performance were also explored in this study. The results showed that the model worked best when using One-hot encoding and truncating the length of the upstream and downstream boundaries within the eccDNA coding region to ≥ 250 bp. The model was trained and tested with different species and tissues to explore its generalizability. The model demonstrated poor generalizability between different species and high generalizability between different tissues of the same species. The results indicated that the model trained with *Arabidopsis* stem tissue was able to predict eccDNA from other *Arabidopsis* tissues and performed better than the model trained with other *Arabidopsis* tissues, which suggested that *Arabidopsis* stem eccDNA contained features of eccDNA from other tissues. Therefore, the eccDNA of other *Arabidopsis* tissues may not need to be detected experimentally and can be predicted directly by the model trained with the *Arabidopsis* stem tissue in future research.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 32070684 and 31571306), and a Project of Shandong Province Higher Educational Program for Introduction and Cultivation of Young Innovative Talents in 2021. Authors would like to thank Supercomputing Center in Shandong Agricultural University for the technical support.

Data availability

eccDNA data is available at the PlantEccDNA (http://123.56.104.85/PlantEccDNA). DeepECD model is available at the GitHub repository of Deep-ECD (https://github.com/HuXiaozz/Deep-ECD).

References

1. Sun H, Lu X, Zou L. 2023. EccBase: A high-quality database for exploration and characterization of extrachromosomal circular DNAs in cancer. Comput Struct Biotechnol J. 21:2591-2601.
2. Gaubatz JW. 1990. Extrachromosomal circular DNAs and genomic sequence plasticity in eukaryotic cells. Mutat Res. 237(5-6):271-292.
3. Kuttler F, Mai S. 2007. Formation of non-random extrachromosomal elements during development, differentiation and oncogenesis. Semin Cancer Biol. 17(1):56-64.
4. Turner KM, Deshpande V, Beyter D, Koga T, Rusert J, Lee C, *et al*. 2017. Extrachromosomal oncogene amplification drives tumor evolution and genetic heterogeneity. Nature. 543(7643):122-125.
5. Zuo S, Yi Y, Wang C, Li X, Zhou M, Peng Q, *et al*. 2022. Extrachromosomal circular DNA (eccDNA): From chaos to function. Front Cell Dev Biol. 9:792555.
6. Wang M, Chen X, Yu F, Ding H, Zhang Y, Wang K. 2021. Extrachromosomal circular DNAs: Origin, formation and emerging function in Cancer. Int J Biol Sci. 17(4):1010-1025.

7. Kumar P, Dillon LW, Shibata Y, Jazaeri AA, Jones DR, Dutta A. 2017. Normal and cancerous tissues release extrachromosomal circular DNA (eccDNA) into the circulation. Mol Cancer Res. 15(9):1197-1205.
8. deCarvalho AC, Kim H, Poisson LM, Winn ME, Claudius Mueller, Cherba D, *et al*. 2016. Extrachromosal DNA elements can drive disease evolution in glioblastoma.
9. Zhu J, Zhang F, Du M, Zhang P, Fu S, Wang L. 2017. Molecular characterization of cell-free eccDNAs in human plasma. Sci Rep.7(1):10968.
10. Zhang P, Mbodj A, Soundiramourtty A, Llauro C, Ghesquière A, Ingouff M, *et al*. 2023. Extrachromosomal circular DNA and structural variants highlight genome instability in *Arabidopsis* epigenetic mutants. Nat Commun. 14(1):5236.
11. Pennisi E. 2017. Unlocking a key to maize's amazing success. Science. 357(6348):240.
12. Møller HD, Parsons L, Jørgensen TS, Botstein D, Regenberg B. 2015. Extrachromosomal circular DNA is common in yeast. Proc Natl Acad Sci U S A. 112(24):E3114-E3122.
13. Cohen S, Méchali M. 2002. Formation of extrachromosomal circles from telomeric DNA in *Xenopus laevis*. EMBO Rep. 3(12):1168-1174.
14. LeCun Y, Bengio Y, Hinton G. 2015. Deep learning. Nature. 521:436–444.
15. Min S, Lee B, Yoon S. 2017. Deep learning in bioinformatics. Brief Bioinform. 18(5):851-869.
16. Oubounyt M, Louadi Z, Tayara H, Chong KT. 2019. DeePromoter: Robust promoter predictor using deep learning. Front Genet. 10:286.
17. Wu P, Nie Z, Huang Z, Zhang X. 2023. CircPCBL: Identification of plant circRNAs with a CNN-BiGRU-GLT model. Plants (Basel). 12(8):1652.
18. Hamamsy T, Morton JT, Blackwell R, Berenberg D, Carriero N, Gligorijevic V, *et al*. 2023. Protein remote homology detection and structural alignment using deep learning. Nat Biotechnol. Epub ahead of print. PMID: 37679542.
19. Abbasi AF, Asim MN, Dengel A, Ahmed S, 2023. iLEC-DNA: identifying long extra-chromosomal circular DNA by fusing sequence-derived features of physicochemical properties and nucleotide distribution patterns. bioRxiv Preprint. https://www.biorxiv.org/content/10.1101/2023.09.01.555875v1.
20. Chang KL, Chen JH, Lin TC, Leu JY, Kao CF, Wong JY, *et al*. 2023. Short human eccDNAs are predictable from sequences. Brief Bioinform. 24(3):bbad147.
21. Wang K, Tian H, Wang L, Wang L, Tan Y, Zhang Z, *et al*. 2021. Deciphering extrachromosomal circular DNA in *Arabidopsis*. Comput Struct Biotechnol J. 19:1176-1183.
22. Lanciano S, Carpentier MC, Llauro C, Jobet E, Robakowska-Hyzorek D, Lasserre E, *et al*. 2017. Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. PLoS Genet. 13(2):e1006630.
23. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 26(6):841-842.
24. Thorvaldsdóttir H, Robinson JT, Mesirov JP. 2013. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 14(2):178-192.
25. Salem M, Arshadi AK, Yuan JS. 2022. AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning. BMC Bioinformatics. 23(1):389.
26. Ullah W, Muhammad K, Haq IU, Ullah A, Khattak S, Sajjad M. 2021. Splicing sites prediction of human genome using machine learning techniques. Multimed Tools Appl. 80:30439-30460.
27. Zhang Y, Hamada M. 2018. DeepM6ASeq: prediction and characterization of m6A-containing sequences using deep learning. BMC Bioinformatics. 19(Suppl 19):524.
28. Lee D, Gorkin DU, Baker M, Strober BJ, Asoni AL, McCallion AS, *et al*. 2015. A method to predict the impact of regulatory variants from DNA sequence. Nat Genet. 47(8):955-961.
29. Krützfeldt LM, Schubach M, Kircher M. 2020. The impact of different negative training data on regulatory sequence predictions. PLoS One. 15(12):e0237412.
30. Jiang M, Anderson J, Gillespie J, Mayne M. 2008. uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. BMC Bioinformatics. 9:192.
31. McCallum A, Nigam K. 1998. A comparison of event models for naive bayes text classification. AAAI Conference on Artificial Intelligence.
32. Breiman L, Friedman JH, Olshen RA, Stone CJ. 1984. Classification and regression trees (cart). Biometrics. 40(3):358.
33. Breiman L. 2001. Random forests. Machine Learning. 45:5–32.
34. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, *et al*. 2011. Scikit-learn: machine learning in python. Journal of Machine Learning Research. 12:2825-2830.
35. Yu HF, Huang FL, Lin CJ. 2011. Dual coordinate descent methods for logistic regression and maximum entropy models. Mach Learn. 85:41–75.
36. Hochreiter S, Schmidhuber J. 1997. Long short-term memory. Neural Computation. 9(8):1735-1780.
37. Chung J, Gülçehre Ç, Cho K, Bengio Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. NIPS 2014 Workshop on Deep Learning.
38. Kim Y. 2014. Convolutional neural networks for sentence classification. conference on empirical methods in natural language processing. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP).
39. Li H, Hu J, Sun X. 2023. Comprehensive evaluation of gene sequence encoding methods in deep learning. https://api.semanticscholar.org/CorpusID:257261909.