Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on Structural Variation Detection Methods of Wheat Genome Based on Deep Learning

Haiping Shi[1,✉], Yanling Li[1], Zijing Dong[1], Yuhong Li[2], Fernando Bacao[3]

[1] *College of Information and Management Science, Henan Agricultural University, Zhengzhou, Henan, 450002, China*

[2] *School of Electrical and Electronic Engineering, Nanyang Technological University, 50 Nanyang Ave, 639798, Singapore*

[3] *NOVA Information Management School (NOVA lMS), Campus de Campolide, Universidade Nova de Lisboa, Lisboa, 1070-312, Portugal*

ABSTRACT

Due to the complexity of genome structure and technical conditions, wheat genome structure variation has not yet been comprehensively and accurately detected and evaluated for genetic effects. The aim of this study is to construct a method based on deep learning algorithm to accurately detect genomic structure variation in wheat. The method converts genomic data into image form by genomic structure variation image generation algorithm. A gene structure variation prediction model is constructed based on deep learning, and efficient and accurate structure variation prediction is realized by automatically extracting and analyzing the variation features in the image. The experimental results show that this method has better detection performance than other structural variation detection methods based on third-generation sequencing data, especially in the structural variation detection of the "Sequencing and Assembly of Spring Wheat Genome in China" project, and the accuracy, precision, and recall rate of this method are all over 90%. This study provides a novel deep learning framework for efficiently detecting structural variants in the wheat genome, and provides powerful technical support for genetic improvement and breeding research of wheat.

Keywords: wheat genes, structural variation, deep learning, image generation algorithm, variation prediction

1. Introduction

Wheat crops, as staple food, feed and energy substances, mainly include common wheat, rye, barley, barley, oats and small rye species [1]. To realize the modernization of agriculture, breeding is the foundation [2]. Wheat, as one of the major food crops in the world today, its breeding is even more important to ensure food security as well as to cope with future climate change [3-4]. Studying the

✉ Corresponding author.
E-mail address: 15137115925@163.com (H. Shi).
Received 05 March 2024; Revised 10 May 2024; Accepted 05 December 2024; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-394

genomic content of wheat crops and utilizing biological tools such as gene editing and transgenics will accelerate the level and process of modern breeding, which is very important for maintaining food security [5-7].

Molecular selection breeding, characterized by the widespread use of single nucleotide markers, has achieved great success during the last decade [8]. However, it has been shown that structural variation as a large-scale genetic variation also plays an important role in phenotypic variation of crops [9-10]. Structural variants (SVs) are widely present in plant genomes and are important factors influencing phenotypic diversity [11]. Due to the complexity of genome structure and technical constraints, structural variants in wheat genome have not been comprehensively and accurately detected and evaluated for genetic effects, and it is urgent to propose effective techniques to accurately detect structural variants in wheat genome [12-14].

In order to analyze the genomes of crops, gene sequencing assembly of relevant crops and sketching of their genomes help to understand the evolutionary pattern of crop genomes and provide important reference information for in-depth study of crop genomic information. Gabur, I. et al. found a strong association between genomic structural variation and crop phenotypic traits by evaluating high throughput methods used to measure genomic structural variation in crop populations, which can be effective in improving resilience and sustainability in order to crop prisons by looking at the pan-genomic diversity of crops [15]. Schiessl, S. V. et al. described the main forms of structural variation in polyploid crop genomes and analyzed them using assays such as optical mapping and long-read sequencing, which can fully understand the relationship between the types of structural variation and crop phenotypes, and provide an optimized pathway for future genetic improvement of polyploid crops [16]. Yuan, Y. et al. pointed out that low-resolution and inefficient gene detection techniques severely limit the understanding of structural variation in plants, so the introduction of structural variation detection methods with greater resolution and accuracy would be beneficial in exploring the diversity of heritable phenotypes within and among species [17].

Some scholars have examined the application of wheat genomic information in wheat breeding. Zhang, Z. et al. sequenced the wheat genome using the PacBio high fidelity (HiFi) sequencing method for detecting structural variations in wheat gene sequences, which played an important role in discovering structurally variable sequences in plants and deciphering their corresponding biological functions [18]. Zhao, J. et al. showed that detecting structural variation in wheat crop genomes plays an important role in understanding the genetic contribution and plasticity of drought resistance in wheat, and that genetic improvement of agronomic traits under drought stress can be enhanced by marker-assisted selection (MAS) breeding based on detecting structural variation in drought tolerance in wheat [19]. Jiao, C. et al. found a large number of structural rearrangements by comparatively analyzing wheat genomes from Chinese breeding history, suggesting that this is a rapid genome evolution caused by wheat's adaptation to its environment, and that analyzing this structural variation could provide a powerful platform for future genome-assisted breeding of wheat [20].

However, with the development of sequencing technology and the explosion of information, more attention has been paid to multiple genomes. Multiple genomes allow for a clearer understanding of the abundance of variation and its impact on modern wheat breeding, and the study of genomics has stepped into the pan-genomic era. Wheat, as a polyploid species, can provide theoretical support for future breeding by constructing deep learning-based pan-genomic structural variation maps to discover structural variation among varieties.

Hexaploid bread wheat (Triticum aestivum L., AABBDD) is one of the world's important food crops, and its excellent yield and adaptability make it a key guarantee for global food security. The genomic structure of bread wheat is complex, containing two different genomes (A, B, D) from three different ancestral species, forming the AABBDD genomic composition. This complex genomic structure endows bread wheat with high genetic diversity and rich adaptability, but also leads to the complexity of genomic rearrangements and gene expression regulation. In the face of challenges posed by global

population growth and climate change, increasing the yield and stress resistance of wheat has become a core objective of agricultural research. In this context, structural variants (SVs) in the genome play a crucial role in the evolution, adaptability, and trait formation of wheat. Structural variation refers to variations such as insertion (INS), deletion (DEL), inversion (INV), duplication (DUP), and translocation (TL) that occur in segments longer than 50 bp in the genome, as shown in Figure 1.

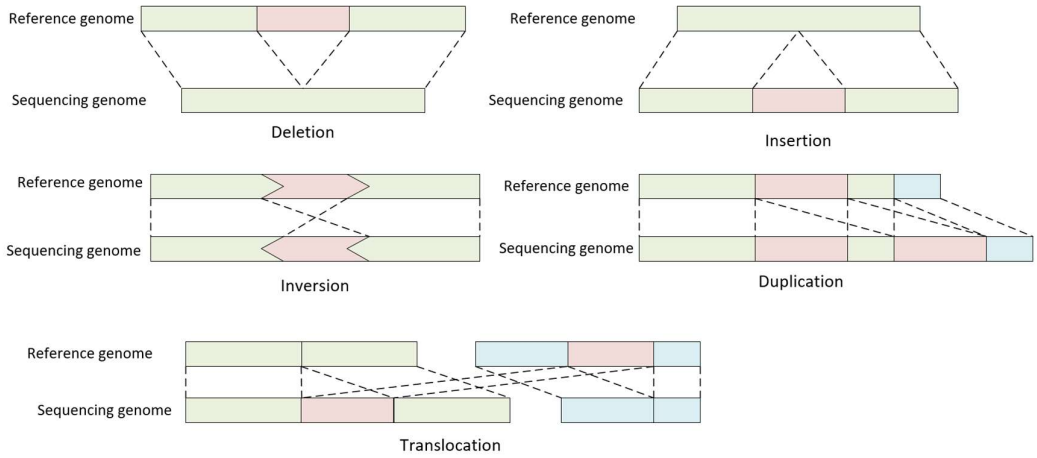


Fig. 1. Types of Structural Variation in the Genome

In the study of structural variation detection in the wheat genome, this paper proposes an innovative method based on deep learning to predict two common and frequently occurring structural variations: deletions and tandem repeats, addressing the limitations of traditional methods in data reading difficulties and low prediction accuracy. This method consists of two core steps: first, using a genome structural variation image generation algorithm to convert genomic data into image form, thereby improving data processing efficiency; Secondly, a deep learning-based model for predicting gene structural variations is constructed, which achieves efficient and accurate structural variation prediction through the automatic extraction and analysis of variation features in images. This method can overcome the shortcomings of traditional approaches, providing a new perspective and tool for the study of structural variations in the wheat genome.

2. Proposed approach

This chapter introduces the image generation algorithm in the detection methods of structural variations in the wheat genome based on deep learning. The algorithm encodes three types of candidate variant interval data—namely, read depth (Read Depth, RD), inconsistent read segments (Discount Read Pair, DRP), and split read segments (Split Read, SR)—using the RGB color model to generate structural variation images. In this way, relevant information about structural variations is intuitively presented, and the output image can clearly display the characteristic information of the variant regions. The workflow of the algorithm is shown in Figure 2:

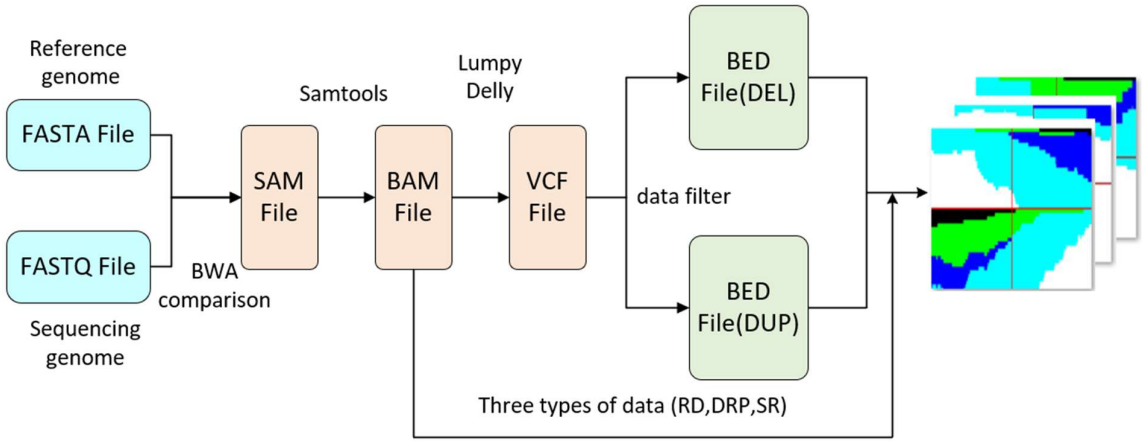


Fig. 2. Flowchart of structural variation image generation in genomes

The core task is to effectively express the representative features of structural variations within a limited-size image. To achieve this goal, it is first necessary to analyze the feature performance of three data types: RD, DRP, and SR, under conditions of genomic deletion and tandem repeat variations, and then encode them into a three-channel image tensor. The image generation process includes steps such as data analysis and extraction, drawing structural variation images, and stitching them together. By expressing the gene sequence alignment information in the form of images, not only has the complexity of extracting variant information from BAM data been reduced, but it has also provided an intuitive genomic input for deep learning models, thereby improving the accuracy of structural variation prediction, as shown in Table 1.

Table 1. Data types and principles used for structural variation

data types	Principle
Read Depth (RD)	The sequencing depth of deletion regions is relatively low, while the sequencing depth of insertion regions is relatively high.
Inconsistent read segments (Discount Read Pair,DPR)	Using the distribution of insert fragment lengths from paired-end sequencing to predict variants, suitable for predicting medium-length insertions and deletions.
Split read segments (Split Read,SR)	During paired-end sequencing, if one end fails to align, there is a possibility of a variant occurring.

3. Data preprocessing

The sequencing data used in this study is sourced from the 'Chinese Spring Genome Sequencing and Assembly' project, and the data can be accessed via the following link: https://www.ebi.ac.uk/ena/browser/view/PRJNA392179. The reference genome selected is the latest version of the Chinese Spring reference genome sequence released by the International Wheat Genome Sequencing Consortium (IWGSC), IWGSC RefSeq v2.1. The FASTQ files generated by the sequencer were aligned to the reference genome (iwgsc_refseqv2.1_assembly.fa) using BWA, from which SAM files containing variant information were obtained. Subsequently, the SAM file is converted to a binary format BAM file using the samtools tool. After preprocessing the BAM file and combining read depth (RD), duplicate read pairs (DRP), and splice site read (SR) data, structural variation images are generated. These images are then used as input data for the deep learning model to perform variant prediction.

The BAM file contains sequence alignment positions and quality information, allowing for precise localization to the reference genome. However, the BAM file itself cannot directly reveal variants; therefore, it is necessary to extract potential variant sites and convert them to VCF format for annotation, classification, and comparison. The VCF file records information such as the location, type, and genotype of variant sites, providing convenience for subsequent analysis. However, the VCF file lacks detailed region annotations, so we convert it to BED format to facilitate the classification, annotation, and comparison of genomic regions, supporting structural variation analysis.

In the algorithm module of this paper, the converted VCF file generates a BED file containing candidate variant regions. Considering that predicted structural variations are typically greater than 50 bp, the algorithm sets a threshold of 50 bp to filter out variations smaller than this threshold, thereby ensuring that the generated BED file contains only valid structural variation regions.

4. Image generation strategy

4.1. Image encoding method

Genomic structural variations include insertions, deletions, inversions, and tandem repeats, among which deletions and tandem repeats are the most common and easily detectable variation types through sequencing technologies. Therefore, this paper focuses on predicting these two types of variations to improve the accuracy and reliability of predictions.

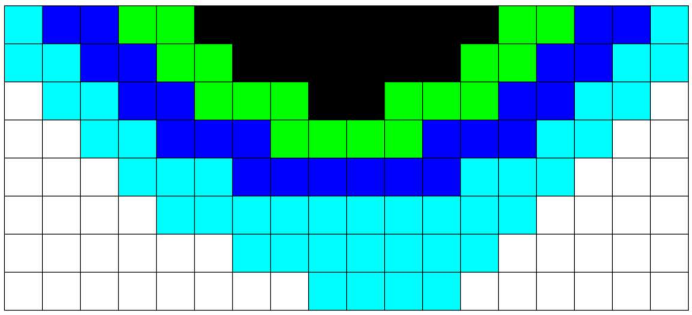
The principle of generating structural variation images is to convert the read fragments of three data types surrounding the candidate variant regions into three-dimensional tensor images. The main objective is to map the alignment information from the BAM file into images, thereby displaying the distribution of RD, DRP, and SR data types within the images. The key challenge lies in how to effectively integrate these three data types into the images. The prediction target of this paper is structural variations exceeding 50 bp, with a focus on the abnormal data alignment of large-scale fragments. In terms of image generation design, the variant image generation algorithm is based on the RGB color model, encoding the RD, DRP, and SR data types with different colors, as described in the specific image generation algorithm below.

1) The R channel is used to represent read depth (RD) data, with the channel value set to a. If the base at that position is covered, the value a of the R channel is 1; otherwise, the value is 255. Therefore, by observing the image of the R channel, the overall trend of read coverage depth in the candidate structural variation loci can be intuitively displayed.

2) The G channel is used to represent inconsistent read segment pairs (DRP) data, with the channel value set to b. If the base at that position is covered and the data type is DRP, then the value b of the G channel is 1; otherwise, the value is 255. Therefore, examining the image of the G channel can clearly reflect the quantity and distribution of the qualifying abnormal alignments of the DRP data type within the image area.

3) The B channel is used to represent split read (SR) data, with the channel value set to c. If the base at this position is covered and the data type is SR, then the value c of the B channel is 1; otherwise, the value is 255. Therefore, examining the image of the B channel can intuitively display the quantity and distribution of SR data types that meet the criteria within the image area.

The mapping of the three data types is shown in Figure 3. The BAM file serves as the canvas for the image, while the data types of each candidate variant interval are represented in the image with corresponding colors through the BED file.



Color	Feature Value	Indicate		
		RD	DRP	SR
	(255, 255, 255)	×	×	×
	(1, 255, 255)	√	×	×
	(1, 1, 255)	√	√	×
	(1, 255, 1)	√	×	√
	(1, 1, 1)	√	√	√

Fig. 3. Colors corresponding to the three alignment strategies

The pixel colors in the image represent the base coverage at the corresponding coordinates. White (255,255,255) represents the background color, cyan (1,255,255) indicates that the base at this coordinate has been covered by the RD type; Blue (1,1,255) indicates that the base coverage at this position comes from both RD and DRP types; Green (1,255,1) indicates that the base coverage at this position comes from both RD and SR types; black (1,1,1) indicates that this position meets all features. The default background color of the image is white (255,255,255); when selecting images, ensure that the size of each image remains consistent.

4.2. Image coverage range

Structural variations typically have a large coverage range, which can vary from 50 bp to several thousand bp, and even for the same type of variation, there may be significant differences in coverage. For each candidate variant interval, the specific positions of the left and right breakpoints can be calculated using formulas (1) and (2):

$$Left = start - 0.1 \times candidate\_len \tag{1}$$

$$Right = end + 0.1 \times candidate\_len \tag{2}$$

Here, Left and Right represent the positions of the left and right breakpoints of the structural variation image, respectively, and the final range of the generated structural variation image is (Left, Right). In this process, start and end represent the starting positions of the candidate variant intervals in the BED file, while candidate_len is the length of the candidate variant, which is the difference between end and start. To ensure that the integrity of the variant region is not compromised when extracting the variant interval, formulas (1) and (2) specify an increase of 0.1 times candidate_len on both sides, which ensures that the generated structural variation image has a certain degree of fault tolerance, avoiding damage to the structure of the candidate variant interval.

4.3. Image stitching rules

For each candidate variant interval's start and end breakpoints, two images of the regions near the breakpoints are formed, including the breakpoints in the images. The stitching rules for the left and right images can be divided into the following three types: stitching along the channel direction (i.e., 3-channel input tensor), horizontal stitching along the x-axis, and vertical stitching along the y-axis. To adapt to the model's input format and avoid forced compression of the sequence range covered by the images, we decided to adopt the last stitching method. When performing left-right flip augmentation based on this stitching method, the positions of the upper and lower images must be swapped simultaneously. The final output is a rectangular image of 224×224 pixels, as shown in Figure 4.

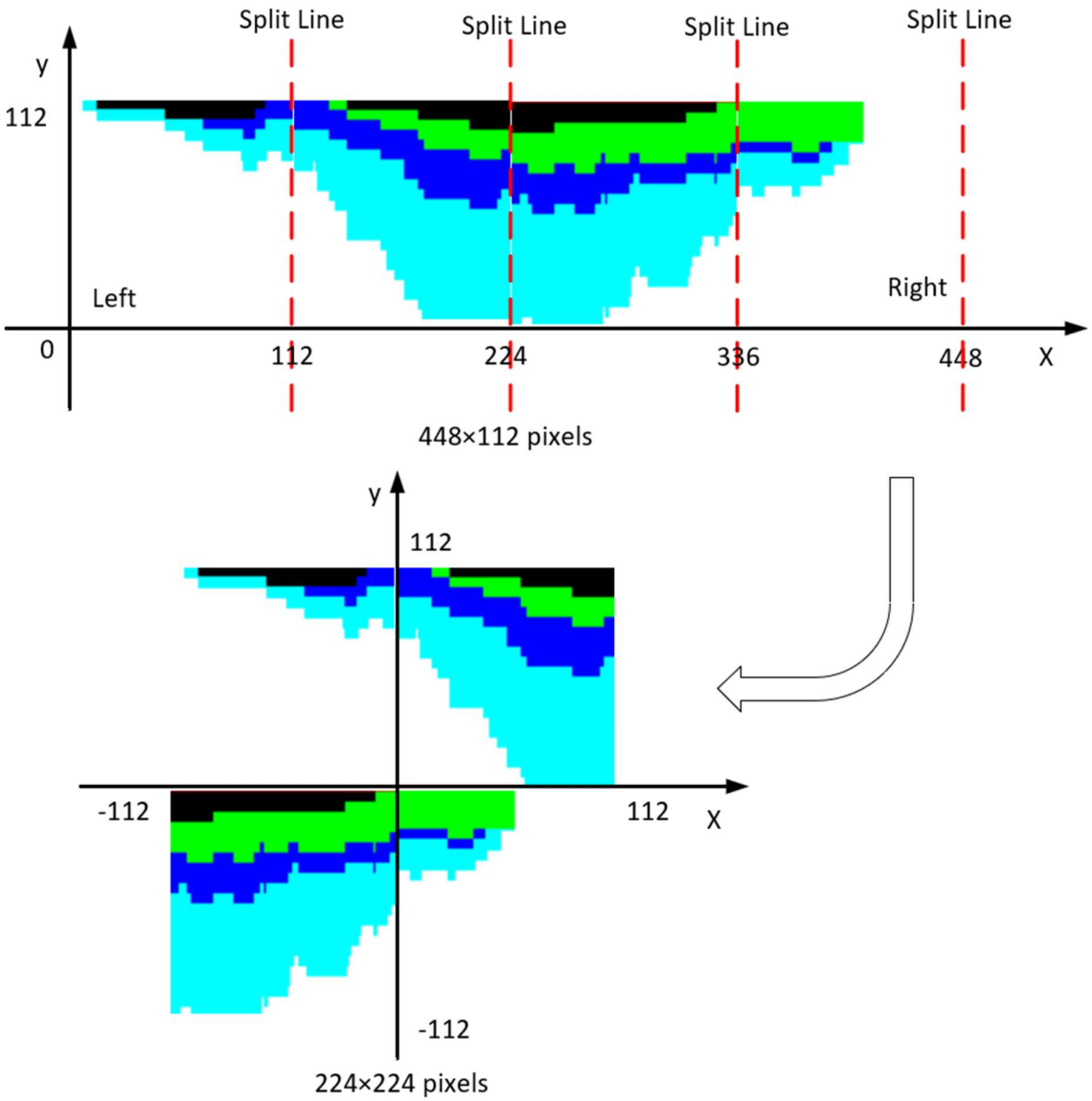


Fig. 4. Structural Variation Image Stitching Strategy

From the following Figure 5, it can be seen that the variation characteristics of DUP are concentrated in quadrants 1 and 3, with a larger coverage area in cyan, blue, and black. In contrast, the variation characteristics of DEL are concentrated in quadrants 2 and 4, with a smaller coverage area in cyan, blue, and black.

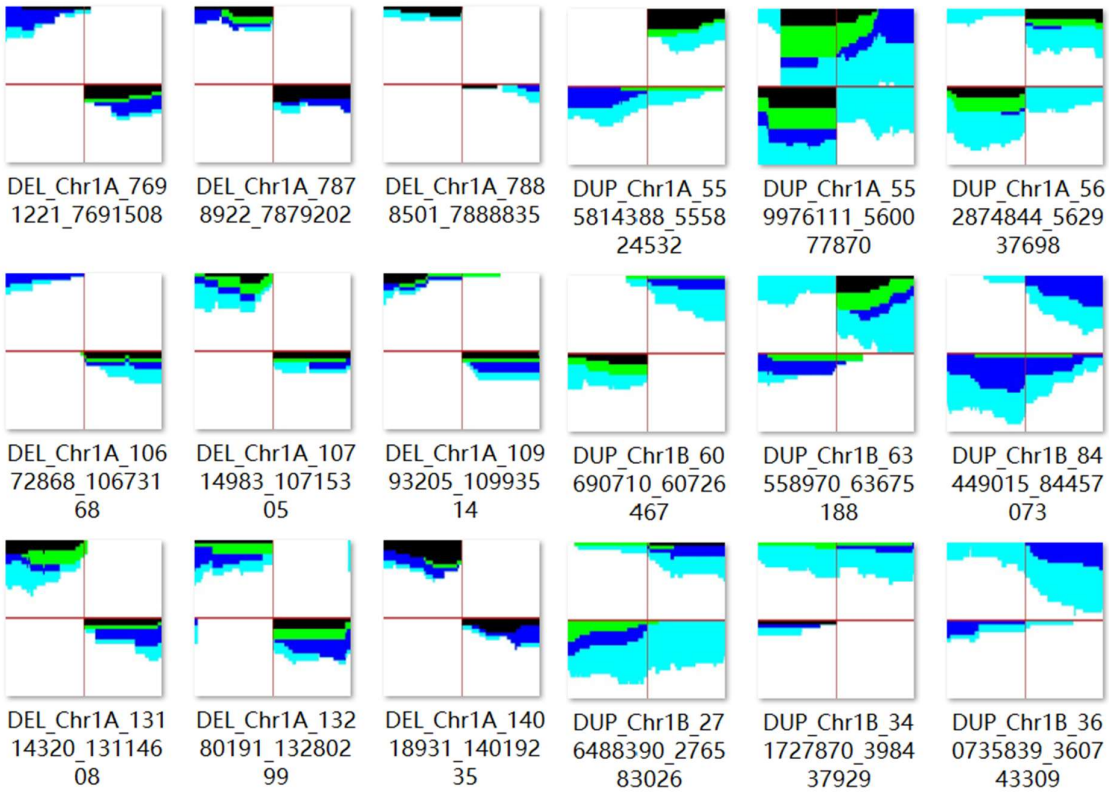


Fig. 5. Display of Image Effects for DEL and DUP Types

5. Structural Variation Prediction Model Based on Deep Learning

5.1. Data Augmentation

Through the above operations, we obtained approximately 500GB of BAM files, and after converting them into VCF files, we generated a BED file containing 86,084 variant information entries. Among them, there are 79,376 deletion variants and 6,708 tandem repeat variants. For each candidate variant interval, we can generate a corresponding structural variation image of the respective type.

When there is a significant difference in the amount of data between the two types of variants, performing a binary classification task may lead to poor classification performance for the smaller category. This is because, during the training process, the model tends to predict samples from the larger category, thereby neglecting samples from the smaller category. To address this issue, the solution proposed in this paper is to enhance the image data through random rotation, Gaussian blur, and adjustments to brightness, contrast, and saturation. The processed results are shown in Figure 6:

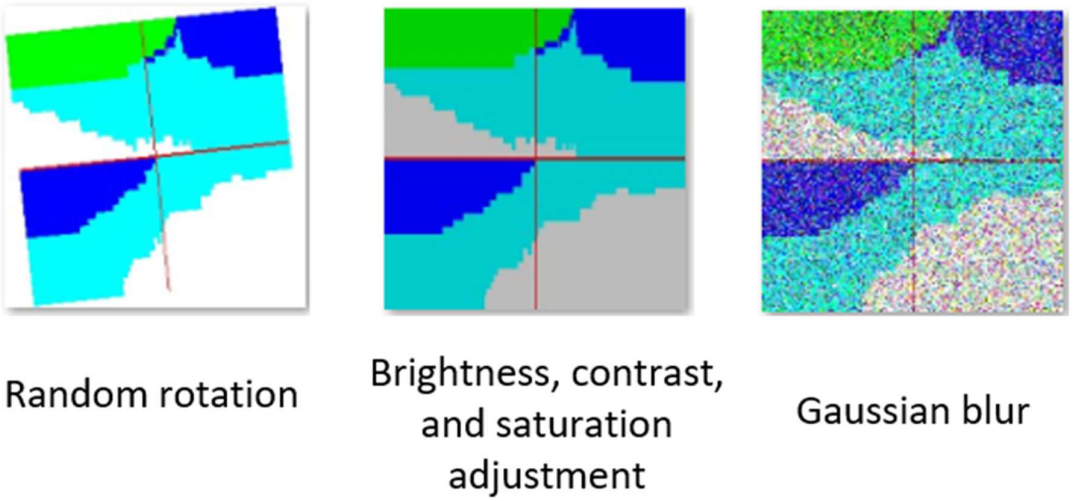


Fig. 6. Demonstration of three image enhancement effects

By applying techniques such as rotation, Gaussian blur, brightness, and contrast adjustment to the training images, the diversity of the dataset has been significantly enhanced. This allows the model to encounter a more diverse set of samples during training, thereby enhancing its generalization ability, particularly in demonstrating greater stability when facing complex scenarios or imperfect data. In addition, to address the issue of class imbalance, data augmentation techniques effectively increased the number of samples in the minority classes, enhancing the model's ability to recognize these classes, which ultimately contributes to improving overall classification performance.

5.2. Experimental process

A total of 12,500 images were used in the experiment, with DEL type and DUP type images each accounting for half. The dataset was randomly divided into 80% training set, 10% validation set, and 10% test set. It is worth noting that only the data in the training set underwent data augmentation, while the data in the validation set and test set did not receive any data augmentation treatment. For the training of the deep learning model, considering factors such as data volume, experimental environment, and computational resources, a batch size of 32 was set, with 50 training epochs, the ReLU activation function was used, the learning rate was set to 0.001, and the Adam optimization algorithm was employed for gradient updates. Apart from the differences in model network parameters and image input preprocessing methods, the training and validation processes are essentially the same.

5.2.1. Image preprocessing and feature extraction. First, the input image is sent to the Patch Partition module for chunk processing. In this module, each image is divided into several adjacent pixel blocks (Patch) of 4x4, and then flattened in the channel direction. Assuming the input image is an RGB three-channel image, each Patch contains 4x4=16 pixels, with each pixel containing R, G, and B values. Therefore, after flattening, the feature dimension of each Patch is 16x3=48. After the Patch Partition module, the shape of the image changes from the original [H, W, 3] to [H/4, W/4, 48], where H and W are the height and width of the image, respectively.

Next, the image undergoes a linear transformation of the channel data for each pixel through the Linear Embedding layer, mapping the 48-dimensional features of each pixel to C dimensions. At this point, the shape of the image changes from [H/4, W/4, 48] to [H/4, W/4, C], where C is a hyperparameter representing the output feature dimension for each Patch. In practice, the Patch Partition and Linear Embedding are accomplished through a convolutional layer, which is similar to the structure of the Embedding layer in traditional Vision Transformers.

5.2.2. Multi-scale construction of feature maps. During the feature extraction process, different sizes of feature maps are progressively constructed through four stages. In Stage 1, the image is first processed through a Linear Embedding layer; In the subsequent three stages, the images at each stage will first undergo downsampling through the Patch Merging layer (which will be detailed later). Then, each stage further extracts features by stacking multiple Swin Transformer Blocks.

It is important to note that two different structures were used when stacking the Swin Transformer Blocks, as shown in Figure 7 (b). The main difference between the two structures is that one uses the W-MSA (Window-based Multi-Head Self Attention) structure, while the other uses the SW-MSA (Shifted Window-based Multi-Head Self Attention) structure. Moreover, these two structures are used in pairs, with the W-MSA structure being used first, followed by the SW-MSA structure. Therefore, when constructing the Swin Transformer, the number of stacked Blocks is usually even (since each pair of Blocks includes one W-MSA and one SW-MSA).

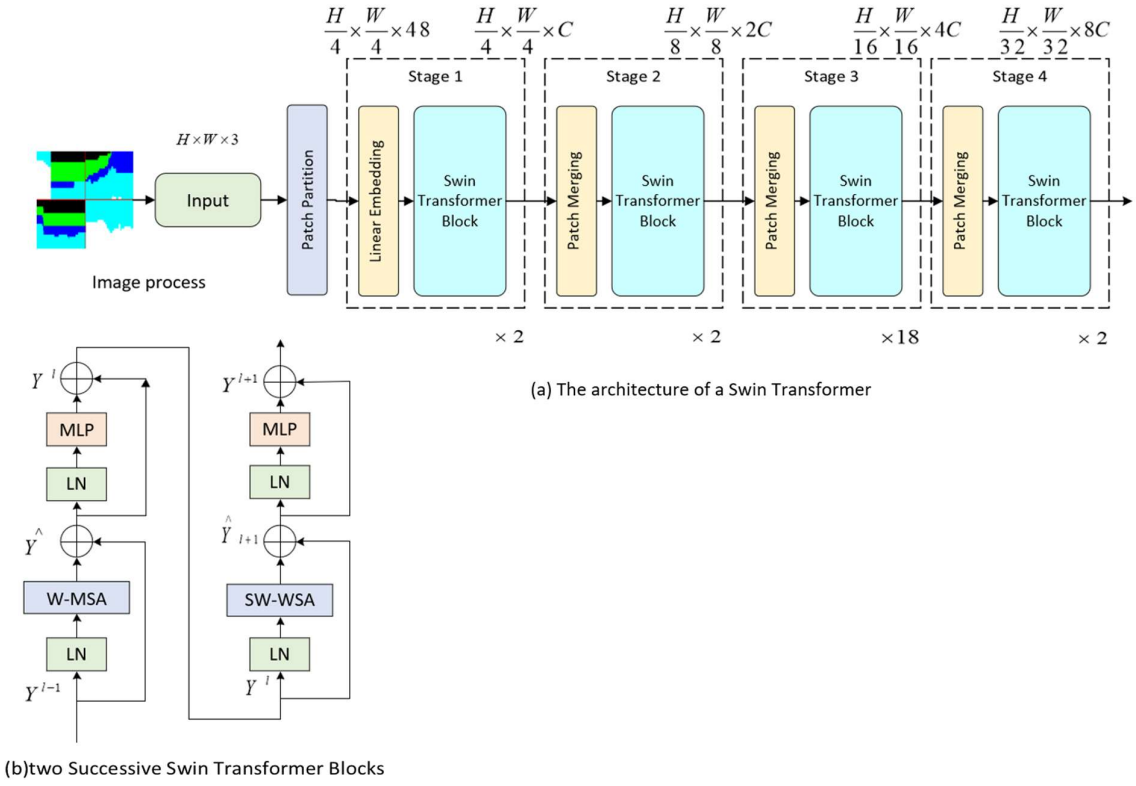


Fig. 7. Swin-Transformer

5.2.3. Classification and Final Output. After completing the multi-layer Transformer Block, the network connects to a Layer Normalization layer, followed by global pooling (Global Pooling) processing. Finally, the final classification results are output through a fully connected layer (Fully Connected Layer). Although these modules are not specifically shown in the diagram, they are indispensable components in the source code implementation.

Since the Swin Transformer introduced a spatial attention mechanism in image classification tasks, demonstrating significant advantages, the corresponding experimental results were obtained after comparing it with six other deep learning models that performed well in image classification—namely, AlexNet, GoogleNet, EfficientNet, ShuffleNet, RegNet, and ResNet; the specific content can be found in Table 2.

Table 2. Experimental results of various models on structural variation images

model	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)

AlexNet	94.32	94.32	95.20	94.37
GoogleNet	96.80	97.14	96.32	96.94
EfficientNet	98.00	98.39	97.60	98.14
ShuffleNet	97.92	98.33	97.60	98.16
RegNet	98.48	98.72	98.24	98.71
ResNet	99.12	98.89	99.36	99.12
Swin-Transformer	99.92	99.84	100	99.84

The results presented in Table 2 indicate that EfficientNet, ShuffleNet, RegNet, ResNet, and Swin-Transformer perform exceptionally well in terms of precision, all achieving over 98%. Specifically, the precision of ResNet and Swin-Transformer is 98.89% and 99.84%, respectively, showing close performance. In terms of recall rate and F1 score, ResNet and Swin-Transformer scored relatively high, with Swin-Transformer demonstrating the best overall performance. Therefore, considering the three metrics of precision, accuracy, and F1 score, Swin-Transformer outperforms other classification models in the final results. As the number of training iterations increases, the trend of loss for the Swin-Transformer on the training set and validation set is shown in Figure 8~ Figure 9, demonstrating the training effectiveness of this deep learning model in the task of structural variation image classification.

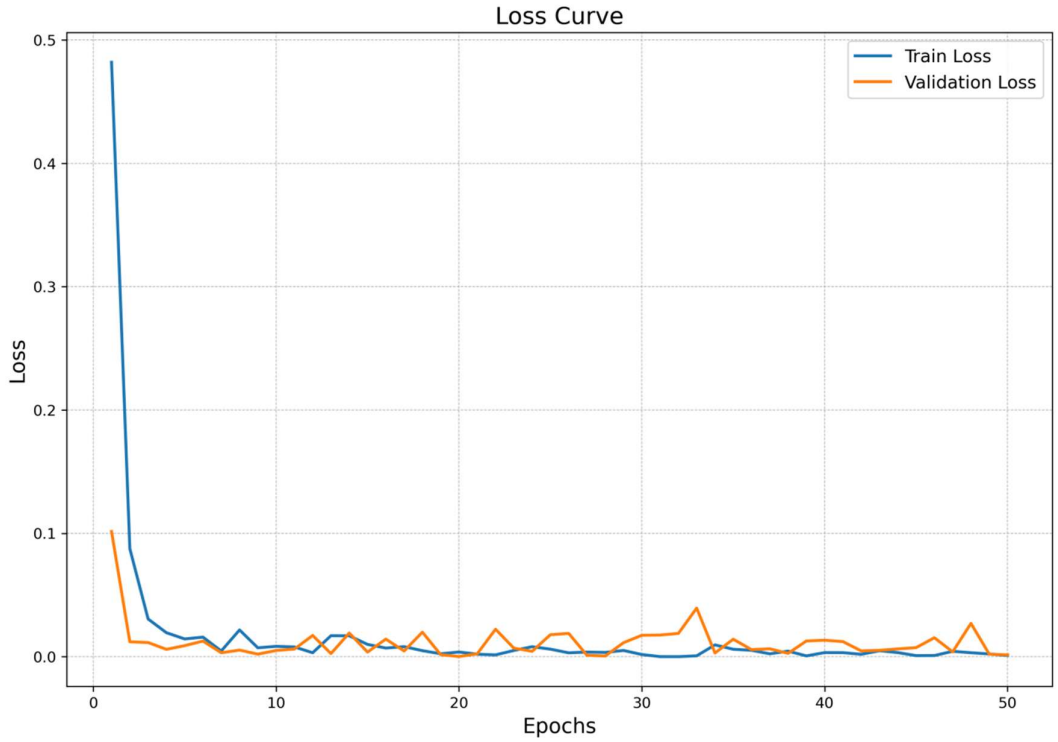


Fig. 8. Loss rate of the Swin-Transformer model as a function of Epoch

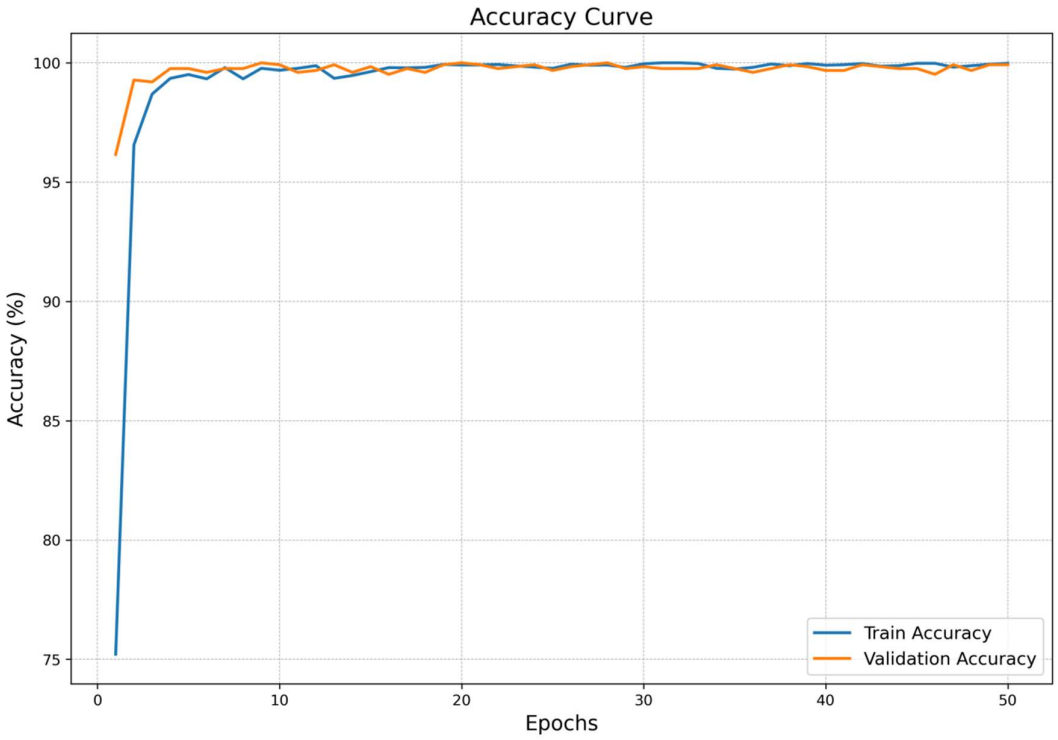


Fig. 9. Accuracy of the Swin-Transformer model during training as a function of Epoch

From the loss rate curve in Figure 8 ~ Figure 9, it can be observed that the loss rates for the training set and validation set began to converge after the 8th round of training, and by the 50th round, the loss rate had stabilized.

5.3. Model training results display

Figure 10 shows the performance of the optimal model obtained from the 8th round of training on the test set. From the figure, it can be seen that all models' ROC curves and AUC values perform excellently, with the ROC curves overall approaching the ideal coordinate point (FPR=0, TPR=1), indicating that the models have a very strong ability to distinguish between positive and negative samples. Meanwhile, the AUC values under the ROC curves of each model are also close to 1. The confusion matrix in Figure 11 displays the classification results of the Swin-Transformer model on the test set, where the true positives (TP) are 625, false negatives (FN) are 0, false positives (FP) are 1, and true negatives (TN) are 624. These data indicate that the Swin-Transformer model performs excellently in classification tasks, with almost no misclassifications and extremely high accuracy. This further validates that the Swin-Transformer model outperforms other comparative models in overall performance, especially achieving outstanding results in recall rate and precision. Therefore, the Swin-Transformer model is considered the best model in this study.

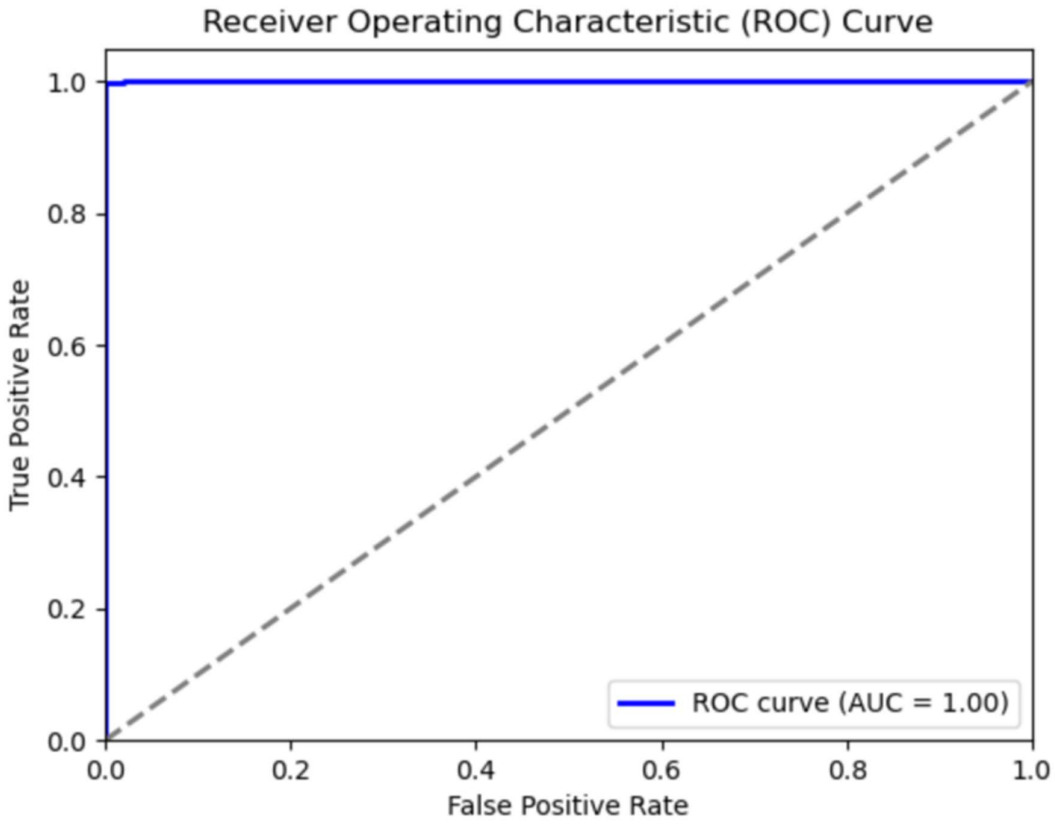


Fig. 10. ROC curve of the Swin-Transformer test set

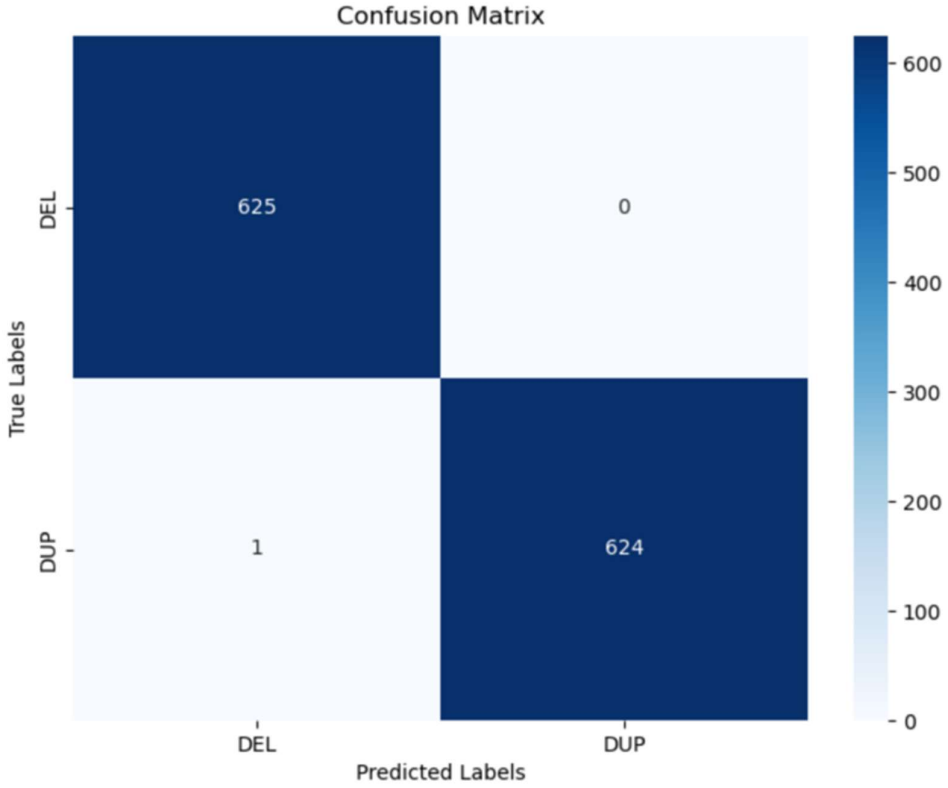
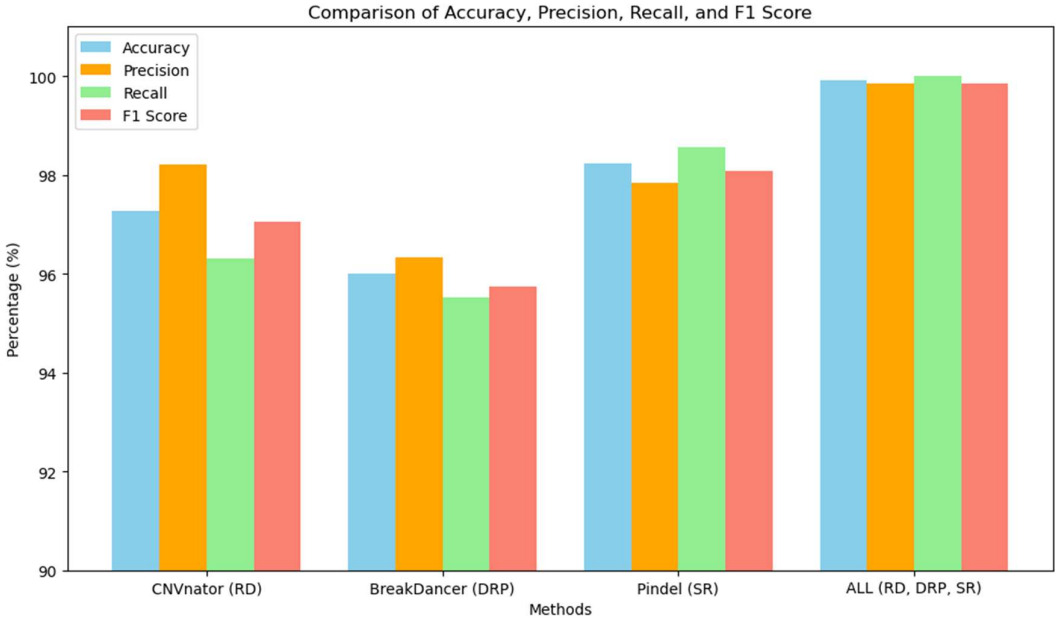


Fig. 11. Confusion matrix of the Swin-Transformer test set

5.4. Comparative experiments

In order to explore the importance of RD, DRP, and SR data types in the classification of structural variation images and whether they are key factors in improving classification performance, this experiment combined three algorithms based on different data types: CNVnator, BreakDancer, and Pindel. In the experiment, we used a fixed network model—Swin-Transformer, and the preprocessing procedures for the training and validation datasets, as well as the network parameter settings, were consistent with those in the previous section. The original dataset used in the comparative experiments of this subsection comes from the 'Chinese Spring Genome Sequencing and Assembly' project (PRJNA392179). The experimental results are shown in Figure 12.



Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
CNVnator (RD)	97.28	98.21	96.32	97.06
BreakDancer (DRP)	96.0	96.34	95.52	95.74
Pindel (SR)	98.24	97.84	98.56	98.08
ALL (RD, DRP, SR)	99.92	99.84	100.0	99.84

Fig. 12. Comparison of accuracy, precision, recall rate, and F1 score

Figure 12 shows the comparison results between the prediction algorithm that combines three types of feature information and the prediction algorithm that uses only a single type of data. The experimental results indicate that when using the DRD data information extracted solely by the BreakDancer algorithm, the prediction performance is significantly inferior to that of the CNVnator algorithm using RD data and the Pindel algorithm using SR data. However, the algorithm presented in image format, which integrates RD, DRP, and SR data, demonstrates the best predictive performance. The comparative experiments above indicate that the algorithm combining these three types of data achieves optimal performance in prediction accuracy, thereby validating the improvements proposed in this paper. In summary, the framework based on sequence-to-image transformation and deep learning methods proposed in this paper outperforms traditional predictive algorithms across multiple metrics. The image encoding scheme that combines RD, DRP, and SR data is both scientifically effective and significantly enhances the classification performance of the model.

6. Conclusion

Wheat gene structure variation detection has been difficult to obtain ideal detection results due to the

diversity of structural variations and complexity of their causes. In this experiment, the deep learning method is introduced to convert the detection of structural variants in wheat genes into an image semantic classification problem in the field of computer vision. And innovative attempts to introduce the image target segmentation method to the structural variation type recognition, after the test shows that all can get high accuracy rate of the gene variation filtering effect, to provide a basis for the subsequent work. However, the current research method can only be used to detect these two types of structural variants, and there are other types of structural variants that can not be detected by the existing algorithms for the time being, such as inversion variants of gene fragments, duplication variants of gene fragments, etc. At the same time, deep learning can be used to recognize structural variants in gene fragments. At the same time, deep learning in the detection of genetic variation, the first feature extraction, followed by the use of neural networks to detect the variant region, which takes a long time. In the future work, we will conduct corresponding research on the detection of other kinds of structural variants, and improve the detection efficiency and shorten the detection time.

Acknowledgements

1) Research on common key technologies of new germplasm resources creation based on artificial intelligence. Project type: Key Research and Development Special Project of Henan Province (231111110100)
2) Research and development and application of information service platform of grain crop germplasm resources, Henan Provincial central government to guide local science and technology development fund (Z20231811005).
3) 2024 Provincial Science and Technology RESEARCH and development Plan Joint Fund (Application research category): wheat germplasm recommendation model study based on FgFisNe network (242103810028).

References

[1] International Wheat Genome Sequencing Consortium (IWGSC), Appels, R., Eversole, K., Stein, N., Feuillet, C., Keller, B., ... & Singh, N. K. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science, 361(6403), eaar7191.

[2] Rasheed, A., & Xia, X. (2019). From markers to genome-based breeding in wheat. Theoretical and Applied Genetics, 132, 767-784.

[3] Uauy, C. (2017). Wheat genomics comes of age. Current opinion in plant biology, 36, 142-148.

[4] Rasheed, A., Mujeeb-Kazi, A., Ogbonnaya, F. C., He, Z., & Rajaram, S. (2018). Wheat genetic resources in the post-genomics era: promise and challenges. Annals of botany, 121(4), 603-616.

[5] Si, Y., Zhang, H., Ma, S., Zheng, S., Niu, J., Tian, S., ... & Li, M. (2025). Genomic structural variation in an alpha/beta hydrolase triggers hybrid necrosis in wheat. Nature Communications, 16(1), 2655.

[6] Xie, Y., Ravet, K., & Pearce, S. (2021). Extensive structural variation in the Bowman-Birk inhibitor family in common wheat (Triticum aestivum L.). BMC genomics, 22, 1-21.

[7] Taagen, E., Tanaka, J., Gul, A., & Sorrells, M. E. (2021). Positional‐based cloning ‘fail‐safe’ approach is overpowered by wheat chromosome structural variation. The Plant Genome, 14(2), e20106.

[8] Song, L., Wang, R., Yang, X., Zhang, A., & Liu, D. (2023). Molecular markers and their applications in marker-assisted selection (MAS) in bread wheat (Triticum aestivum L.). Agriculture, 13(3), 642.

[9] Gimenez, K., Blanc, P., Argillier, O., Kitt, J., Pierre, J. B., Le Gouis, J., & Paux, E. (2025). Impact of structural variations and genome partitioning on bread wheat hybrid performance. Functional & Integrative Genomics, 25(1), 10.

[10] Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., ... & Clark, M. D. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome research, 27(5), 885-896.

[11] Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., ... & Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. Nature, 588(7837), 277-283.

[12] Badet, T., Fouché, S., Hartmann, F. E., Zala, M., & Croll, D. (2021). Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. Nature communications, 12(1), 3551.

[13] Kong, L., Cheng, H., Zhu, K., & Song, B. (2025). LOGOWheat: deep learning–based prediction of regulatory effects for noncoding variants in wheats. Briefings in Bioinformatics, 26(1), bbae705.

[14] Wang, X., Xuan, H., Evers, B., Shrestha, S., Pless, R., & Poland, J. (2019). High-throughput phenotyping with deep learning gives insight into the genetic architecture of flowering time in wheat. GigaScience, 8(11), giz120.

[15] Gabur, I., Chawla, H. S., Snowdon, R. J., & Parkin, I. A. (2019). Connecting genome structural variation with complex traits in crop plants. Theoretical and applied genetics, 132, 733-750.

[16] Schiessl, S. V., Katche, E., Ihien, E., Chawla, H. S., & Mason, A. S. (2019). The role of genomic structural variation in the genetic improvement of polyploid crops. The Crop Journal, 7(2), 127-140.

[17] Yuan, Y., Bayer, P. E., Batley, J., & Edwards, D. (2021). Current status of structural variation studies in plants. Plant Biotechnology Journal, 19(11), 2153-2163.

[18] Zhang, Z., Zhang, J., Kang, L., Qiu, X., Xu, S., Xu, J., ... & Lu, F. (2024). Structural variation discovery in wheat using PacBio high‐fidelity sequencing. The Plant Journal, 120(2), 687-698.

[19] Zhao, J., Li, X., Qiao, L., Zheng, X., Wu, B., Guo, M., ... & Zheng, J. (2023). Identification of structural variations related to drought tolerance in wheat (Triticum aestivum L.). Theoretical and Applied Genetics, 136(3), 37.

[20] Jiao, C., Xie, X., Hao, C., Chen, L., Xie, Y., Garg, V., ... & Zhang, X. (2024). Pan-genome bridges wheat structural variations with habitat and breeding. Nature, 1-10.