

1 Deep learning empowers genomic selection of pest-resistant 2 grapevine

3 Yu Gan^{1,2,†}, Zhenya Liu^{1,2,†}, Fan Zhang^{1,2,†}, Qi Xu^{2,†}, Xu Wang^{1,2}, Hui Xue^{1,2},
4 Xiangnian Su^{1,2}, Wenqi Ma², Qiming Long², Anqi Ma², Guizhou Huang², Wenwen
5 Liu², Xiaodong Xu^{1,2}, Lei Sun³, Yingchun Zhang², Yuting Liu², Xinyue Fang²,
6 Chaochao Li^{1,2}, Xuanwen Yang^{1,2}, Pengcheng Wei¹, Xiucui Fan³, Chuan Zhang⁴,
7 Pengpai Zhang⁶, Chonghuai Liu³, Lianzhu Zhou^{1,2}, Zhiwu Zhang⁵, Cong Tan⁷, Yiwen
8 Wang^{1,2,*}, Zhongjie Liu^{1,8,*}, Yongfeng Zhou^{1,*}

9 1 National Key Laboratory of Tropical Crop Breeding, Tropical Crops Genetic Resources
10 Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China

11 2 National Key Laboratory of Tropical Crop Breeding, Shenzhen Branch, Guangdong
12 Laboratory of Lingnan Modern Agriculture, Key Laboratory of Synthetic Biology, Ministry of
13 Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy
14 of Agricultural Sciences, Shenzhen, China

15 3 Zhengzhou Fruit Research Institute, Chinese Academy of Agricultural Sciences, Zhengzhou,
16 China

17 4 State Key Laboratory of Genetic Improvement and Germplasm Innovation of Crop Resistance
18 in Arid Desert Regions (Preparation), Key Laboratory of Genome Research and Genetic
19 Improvement of Xinjiang Characteristic Fruits and Vegetables, Institute of Horticultural
20 Crops, Xinjiang Academy of Agricultural Sciences, Urumqi, China

21 5 Department of Crop and Soil Sciences, Washington State University, Pullman, WA, USA

22 6 School of Life Sciences, Henan University, Kaifeng 475004, China

23 7 BGI Research, Shenzhen, 518083, China

24 8 Institute of Life and Health, China Resources Research Institute of Science and Technology,
25 Hong Kong, China

26 † These authors contributed equally to this work.

27 * Corresponding authors: zhouyongfeng@caas.cn; liuzhongjie@caas.cn; wangyiwen@caas.cn

28

29 © The Author(s) 2025. Published by Oxford University Press. This is an Open Access
30 article distributed under the terms of the Creative Commons Attribution License
31 <https://creativecommons.org/licenses/by/4.0/>, which permits unrestricted reuse,
32 distribution, and reproduction in any medium, provided the original work is properly
33 cited.

34

Abstract

Crop pests significantly reduce crop yield and threaten global food security. Conventional pest control relies heavily on insecticides, leading to pesticide resistance and ecological concerns. However, crops and their wild relatives exhibit varied levels of pest resistance, suggesting the potential for breeding pest-resistant varieties. This study integrates deep learning (DL)/machine learning (ML) algorithms, plant phenomics, quantitative genetics and transcriptomics to conduct genomic selection (GS) of pest-resistance in grapevine. Building deep convolutional neural networks (DCNN), we accurately assess pest damage on grape leaves, achieving 95.3% classification accuracy (VGG16) and a 0.94 correlation in regression analysis (DCNN-PDS). The pest damage was phenotyped as binary and continuous traits and genome resequencing data from 231 grapevine accessions were combined in a Genome-Wide Association Studies (GWAS), which maps 69 QTLs and 139 candidate genes involved in pest resistance pathways, including jasmonic acid, salicylic acid, and ethylene. Combining this with transcriptome data, we pinpoint specific pest-resistant genes such as *ACA12* and *CRK3*, which are crucial in herbivore responses. Machine learning-based GS demonstrates a high accuracy (95.7%) and a strong correlation (0.90) in predicting pest resistance as binary and continuous traits in grapevine, respectively. In general, our study highlights the power of DL/ML in plant phenomics and genomic selection, facilitating genomic breeding of pest-resistant grapevine.

Key words: Viticulture, smart breeding, AI breeding, GWAS, GS, DCNN, plant phenomics

60 Introduction

61 The adverse impacts of agricultural pests have long posed challenges for agriculture¹.
62 Pests, such as aphids, locusts, and moth flies directly damage crops, resulting in
63 reduced yields and compromised quality². Other pests like scale insects and thrips
64 indirectly harm crops by vectoring plant viruses³. The strong reproductive ability of
65 these pests not only decreases crop production but also threatens food security⁴.
66 Traditionally, farmers have primarily relied on copious applications of pesticides to
67 control crop pests. However, the excessive use of pesticides contaminated the
68 environment and jeopardized human and animal health⁵. Moreover, the rapid
69 evolution of pesticide resistance among insect populations has progressively
70 diminished the efficacy of these chemicals⁶. Thus, new monitoring methods and
71 control strategies are needed to precisely control agricultural pests.

72 Plants and insects have coexisted for over 350 million years⁷. As stationary organisms,
73 plants lack the ability to escape attacks from other organisms, requiring the use of
74 alternative strategies for self-defense. To counter the herbivore attack, plants develop
75 specialized structures or produce secondary metabolites and proteins with toxic,
76 repellent, or antinutritional effects to deter herbivores^{8,9}. Plants defend against
77 herbivores both directly, by altering the preferences, survival and reproductive success
78 of herbivores, and indirectly, by emitting volatile organic compounds (VOC) to attract
79 their natural enemies¹⁰. The defense mechanisms of plants against herbivore attacks
80 involve intricate signal transduction pathways mediated by a network of
81 phytohormones. Plant hormones play a pivotal role in governing plant growth,
82 development, and defense mechanisms¹¹. Several plant hormones play roles in both
83 intra- and inter-plant communication in herbivore-damaged plants. The majority of
84 defense responses against insects are triggered by signal transduction pathways
85 involving jasmonic acid (JA), salicylic acid (SA), and ethylene¹⁰. Distinct sets of
86 genes associated with defense are activated by these pathways in response to injury or
87 pest feeding. These hormones can work independently, synergistically, or
88 antagonistically, depending on the specific attacker.

One major pest posing prominent threats to grape yield and quality is the tobacco cutworm (*Spodoptera litura*, Fab, SLF)¹². In grape cultivation, infestations generally occur between July and September, especially in drought conditions, high humidity, and in proximity to vegetable crops. The larvae of SLF damage the plant mainly by consuming leaf mesophyll tissue, resulting in perforated foliage¹³. Because outbreaks mainly transpire during grape ripening and harvest, extensive feeding can disrupt grape coloring and maturation in the same year. This also impedes nutrient translocation in the following season, leading to poor subsequent bud break and flowering^{14,15}. Although a few grape varieties show resistance to SLF, resistant varieties have not been selected and bred to resist SLF infestations. Therefore, determining more resistance germplasm resources, genetics of plant resistance to SLF and the function of resistance genes have been essential for breeding pest-resistant grapevine.

While the throughput and cost-effectiveness of genotyping have improved considerably¹⁶⁻¹⁸, the measurement of traits of interest remains insufficient in these regards. Most of the resistance traits have been subjected to visual and sensory evaluations by experienced professional breeders, and phenotypic values of the traits are expressed as qualitative categorical scores¹⁹. For example, drought stress level (mild, moderate and severe, three categories)²⁰ and downy mildew level (seven categories)²¹. An empirical assessment based on the sense of the breeder is not sufficient to evaluate the diverse and continuous variations of the plant resistance. In addition, expertise in visual and sensory evaluation can only be obtained after years of training, and increasing the number of specialized breeders is not practical for large scale and low-cost phenotyping. To further improve the accuracy of GWAS and GS for practical breeding, we need to enhance the data quality of the agronomic traits²². Deep learning-based image analysis offers a promising approach to address the limitations of current qualitative evaluation methods.

Deep learning is a subset of machine learning that utilizes artificial neural networks to learn from large volumes of data. Deep neural networks contain multiple hidden

layers between the input and output layers that enable the model to progressively extract features at multiple levels of abstraction²³. This hierarchical learning architecture allows deep learning models to learn complex functions and find intricate patterns in data that most traditional machine learning techniques may fail to detect. In agriculture, deep learning has emerged as a promising approach for improving the detection and classification of pests, diseases, and other anomalies. For example, Mohanty et al. developed a deep convolutional neural network able to detect 14 crop diseases and 26 crop species with an accuracy exceeding 95%²⁴. Sladojevic et al. used deep neural networks for automated identification of plant diseases from leaf images, achieving an accuracy of 96.3% compared to 85.9% with a shallow neural network²⁵. Overall, deep learning allows predictive models for agricultural applications to achieve greater complexity and abstraction by learning directly from raw data, enabling more accurate and nuanced detection capabilities than traditional analytical techniques. The ongoing explosion of agricultural big data presents an ideal opportunity for deep learning techniques to continue advancing in this area.

In this study, we developed deep learning models that combined object detection and pest severity assessment to generate pest resistance phenotype data from 231 grapevine accessions. Using GWAS, we identified QTL regions associated with resistance and identified pest-resistance related candidate genes within these regions. Furthermore, we integrated GWAS with population genetic analysis, transcriptomic assays, comparisons among different *Vitis* populations, and inter-species comparisons to determine which genomic regions contribute to resistance, as well as to identify potential candidate resistance genes within these regions. Finally, we employed machine learning based GS to evaluate different models in breeding programs of pest-resistance grapevine. Overall, our research deepens insight into pest resistance genetics and evolution, promoting advanced genomic breeding strategies for grapevine resistance, and outlines a pathway for potential broader application to other crops.

Results

The performances and cross-validations of different binary-classification models

To build the binary classification models for the categorization of damage levels caused by insects, we employed a set of DCNNs, including AlexNet, VGG16, ResNet50, ResNet101, InceptionV3, and DenseNet121. The model performance was rigorously evaluated through cross-validation (CV) with the aim of identifying the most optimal hyperparameters. After hyperparameters tuning, the best model was then selected based on the results from the test set. The entire CV process utilized a total of 1448 leaf images for training and validation purposes. After choosing the best hyperparameters and training the models for sufficient epochs to ensure convergence, the highest average accuracy was achieved by VGG16 at 90.1%, while the lowest accuracy was observed in ResNet101, scoring 86.9%. Other models fell within the range of 87% to 89% accuracy (Fig. 1A). Correspondingly, VGG16 also exhibited the highest average F1 score, reaching 0.90, while others ranged from 0.86 to 0.89.

Under the condition that all six models are optimized, we further used the test set (362 images) to determine which of the six models is most suitable for our binary classification task. We found that VGG16 continued to perform the best on the test set, achieving an accuracy of 95.3% (Supplementary Table 1). Upon scrutinizing the confusion matrix, it was evident that, the model accurately classified 169 positive and 176 negative samples (Fig. 1B). We then performed a correlation analysis between accuracy, F1 score, and model complexity. This revealed a negative correlation trend between the model's complexity and its performance (Supplementary Fig. 1). VGG16 exhibited the best performance with a relatively low complexity, making it our choice for the binary classification task. The learning curves of VGG16 were shown in Supplementary Fig. 3.

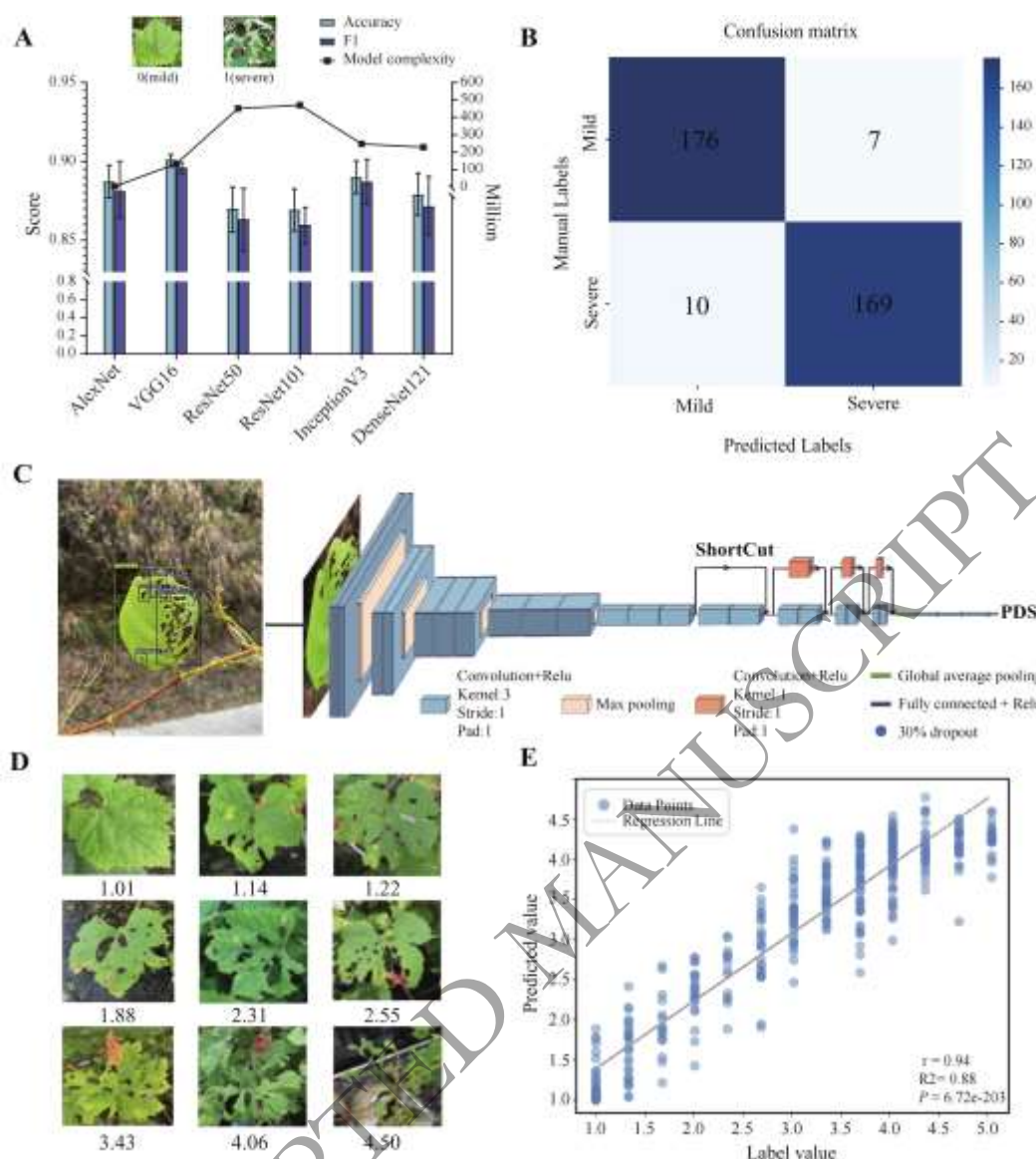


Fig. 1 | The deep learning model (DCNN) for phenotyping pest resistance as binary and continuous traits. A, Performance of six classic convolutional neural networks in CV. The bars correspond to the left Y-axis (accuracy score), while the line is referenced to the right Y-axis (model complexity). **B,** The number of correct recognitions of different categories by VGG16 in the test set. **C,** The model architecture of DCNN-PDS and the workflow for deriving continuous traits from images. The model uses pretrained VGG16 for initial feature extraction, followed by four residual blocks (each consisting of two convolutional layers) to enhance deep feature learning and prevent degradation. A global average pooling (GAP) layer is then applied to reduce the feature dimensions, and three fully connected layers are used for feature fusion. **D,** Examples of the DCNN-PDS prediction results for the different extent damages by pest, and the PDS values are shown at the bottom of each plot. **E,** The correlation between the predicted values provided by DCNN-PDS and the manually labeled values in the test set.

The performance of DCNN-PDS on continuous traits

To further enhance the granularity of damage assessment caused by pests on leaves, we devised and trained a regression model to assign continuous scores ranging from 0-5, simultaneously providing more suitable data for genetic analysis (Fig. 1C). The pest damage regression model underwent CV and was subsequently evaluated on an independent test set to ensure that there was no overfitting. During CV, a dataset of 2,181 images was used to identify the optimal hyperparameters configuration. Through hyperparameters tuning, the model achieved an average mean squared error (MSE) loss of 0.1245 and an average mean absolute error (MAE) loss of 0.2343 in CV. (Supplementary Table 2). After the model was configured, formal training was conducted. The model was initially set to train for 300 epochs. Because of early stopping, the training was halted after 218 epochs (Supplementary Fig. 4). After DCCN-PDS (pest damage score) converged, the MSE loss decreased to 0.0241 on the training set and 0.0866 on the validation set, while the MAE dropped to 0.1061 for the training set and 0.1845 for the validation set.

Before applying the model in practical settings, we captured 439 new images for testing to ensure its generalization and robustness (Supplementary Fig. 5). During the testing phase, the model achieved an MSE of 0.1946 and an MAE of 0.3504 on the test set. The DCCN-PDS prediction example showcases the model's precise quantification of subtle variations in pest damage severity (Fig. 1D). To further analyze the relationship between the predicted values and the manual tag values, a correlation analysis was performed. A significant correlation ($P < 0.001$) between the manual tag values and the predicted values on the test set was observed, as indicated by a Pearson correlation coefficient of 0.94, with a Coefficient of Determination $R^2 = 0.88$ (Fig. 1E).

Phenotypic assessment of grape accessions using binary classifications and regression models

To assess the pest-resistance of grapevine populations, we collected cuttings of 231

accessions from National Grape Germplasm at Zhengzhou in Nov. 2021 and cultivated in greenhouse at the Shenzhen station (Fig. 2A and Supplementary Fig. 6). After the outbreak of SLF in the greenhouse in June-August 2022 (Fig. 2A, see Methods), we randomly take pictures for 6-10 leaves for each accession in September 2022 to represent the pest-resistance of the sample. A population structure analysis was performed on a total of 231 grape varieties, including table grape varieties domesticated from *V. vinifera* ssp. *Sylvestris* (EU-Table), table grape varieties originated from hybridization between *V. vinifera* and *V. labrusca* (EA-Table), and wine-making grape varieties domesticated from *V. vinifera* ssp. *sylvestris* (Wine) (Figure 2 and Supplementary Table 3). The examination utilized both the VGG16 model and our custom-developed model DCNN-PDS, resulting in binary and continuous phenotypes (Fig. 2B). Binary classification results revealed 99 varieties exhibited mild damage by pests and 132 varieties have suffered severe pest damage (Supplementary Fig. 7). The EA-table seems to be the least resistant to pests, with 66 out of 92 accessions experiencing severe infestation (50% of the total severe grape accessions). In contrast, the EU-table appears to have stronger pest resistance, with 46 varieties identified as having mild infestation (46.47% of the total instances of mild damage) (Fig. 2C, D).

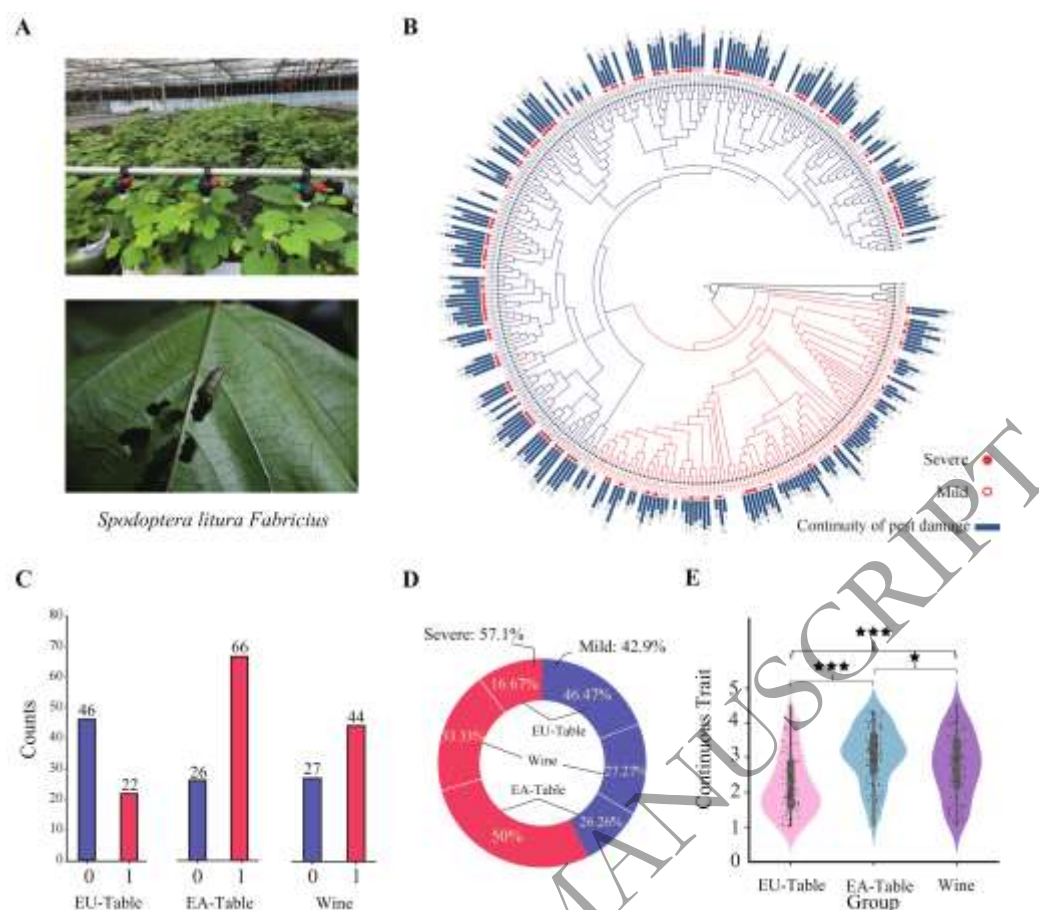


Fig. 2 | The overall phenotype of pest resistance in grapevine populations. **A**, The greenhouse conditions and the experimental setup for SLF infestation. **B**, The evolutionary tree of 323 individuals is divided into two populations: Europe and America, and Eurasia. The bar chart displays continuous trait values for each accession, with solid red dots indicating severity in binary traits. The footrule distance between the species order generated by the phylogenetic tree and the insect resistance order is 19,405, whereas the average distance from 1,000 randomly generated ordering simulations is 23,102, a significant deviation from the observed result (one-sample t-test, $P < 2.2 \times 10^{-16}$). **C**, For the three grape populations based on the classification results identified VGG16, the blue bars correspond to the x-axis value '0', representing healthy samples, while the red bars correspond to the x-axis value '1', representing samples identified as having severe pest damage. **D**, A pie chart depicting the overall binary classification phenotypes, and described the percentage of their respective binary phenotypes. **E**, Violin plots depicting the distribution of continuous phenotypes for the three categories, with significance differences denoted by asterisk (★: $P < 0.05$, ★★★: $P < 0.001$).

DCCN-PDS predicted and scored damage levels in 0-5 for all accessions based on their images (Supplementary Fig. 7). Regression analysis of the leaf images revealed the EU-table group had the lowest mean and median values at 2.22 and 1.94, respectively (Fig. 2E). Normality testing using the Shapiro-Wilk test indicated the EU-wine group followed a normal distribution ($P = 0.082$), while the P values for the other two groups were below 0.05 (Supplementary Fig. 8). Homogeneity of variance among the three groups was supported by a Bartlett test ($P = 0.5542$). Given these results, the Wilcoxon rank sum test was used to detect differences between groups and found significant differences between all three. It was evident that, compared to the EA-table and EA-wine groups, the EU-table exhibited significant ($P < 0.001$) phenotypic distinctions, indicating a generally heightened pest resistance compared to the aforementioned two populations (Fig. 2E). Phylogenetic regression was performed by mapping regression phenotypes onto the phylogenetic tree (Fig. 2B), and correlation and Spearman's footrule distance were calculated. The correlation between regression value order and tree order was 0.2747, with a Spearman's footrule distance of 19405. In contrast, the average correlation of random sequences was $-3.09\text{e-}4$, with an average distance of 23102. Single sample t-tests showed both analyses were significantly non-random ($P < 2.2\text{e-}16$), suggesting a significant association between pest resistance and phylogenetic evolution.

Genome-wide Association Study (GWAS) of binary and continuous pest-resistant traits

To further map the quantitative genetic basis of pest damage, we conducted GWAS for binary traits derived from VGG16 and for continuous traits obtained from DCCN-PDS, respectively (Fig. 3A and Supplementary Fig. 9). This aimed to investigate their ability to efficiently identify genetic loci associated with pest resistance. The analysis revealed 33 significant QTLs associated with binary traits, involving 67 candidate genes, and 36 QTLs associated with continuous traits, encompassing 85 candidate genes (Supplementary Tables 4 and 5). Genes associated with continuous traits, as identified through GO enrichment analysis, were found to be primarily involved in the

calcium ion binding process ($P < 0.05$). In contrast, the genes associated with binary traits did not exhibit significant enrichment in molecular functions, nor did the total 139 non-redundant candidate gene sets (Supplementary Table 6). Notably, all three gene sets detected a cluster related to involvement in protein kinase activity and protein phosphorylation, although the significance level was not sufficient (Supplementary Table 7). We identified *1-aminocyclopropane-1-carboxylate oxidase homolog 1* (ACO, Vitvi008234) and *sterol 3-beta-glucosyltransferase UGT80B1* (TT15, Vitvi022534) on chromosomes 5 and 12, respectively, associated with binary phenotype results, as well as *probable protein phosphatase 2C 14* (PP2C, Vitvi000748) on chromosome 1 linked to continuous phenotype results, all of which have been previously implicated in pest resistance in plants (Fig. 3A). Vitvi008234 and Vitvi022534 are involved in the synthesis of ethylene and flavonoids (Supplementary Table 8), respectively. These two compounds play a significant role in influencing the plant's defense against herbivore feeding^{26,27}. Vitvi000748 encodes a PP2C-type phosphatase (Supplementary Table 9), PP2C has been identified as a negative regulator of protein kinase cascades activated in response to stress²⁸, existing research has indicated that PP2C is involved in the plant's defense response against herbivores²⁹. In addition, we identified a strong signal (Chr16: 26.7M-27.0M) for continuous trait (snp_16_26926513, $P = 1.71e-06$). Plants possessing both pure and mutated loci demonstrate lower scores and diminished pest damage (Fig. 3B). The predominant locus among the three populations is characterized by heterozygous haplotypes, with a higher prevalence of homozygous mutant haplotypes in the EU-Table and an increased frequency of homozygous mutant haplotypes in the EA-Table (Fig. 3C). Additionally, the genome-wide analysis of genetic differentiation (F_{ST}) and genetic diversity (π) indicates an absence of selection events in this region (Fig. 3D). The linkage disequilibrium analysis highlights a significant linkage in the promoter region of *CDPK-RELATED KINASE 3* (CRK3, Vitvi031045) (Fig. 3E). The gene Vitvi031045 encodes a Ca^{2+} /CaM-dependent protein kinase related protein kinase and participate in CaM-mediated signaling pathways³⁰. (Supplementary Table 7). Previous studies have demonstrated that CRK overexpression and knockout mutants exhibits

increased and decreased defense traits against SLF³¹. Furthermore, it was found that the transcription of *CRK3* is positively regulated by JA and abscisic acid (ABA) signals, particularly under herbivory-induced stimulation³¹. It plays a pivotal role in modulating plant defense responses by phosphorylating and activating *WRKY14*, and enhancing the defense response through the upregulation of *PDF1.2*³¹.

ACCEPTED MANUSCRIPT

differences between them (★: $P < 0.05$, ★★: $P < 0.01$, ★★★: $P < 0.001$). C, The number of individuals corresponding to three genotypes in three populations, where 0/0 represents no mutation, 0/1 represents heterozygous mutation, and 1/1 represents homozygous mutation. D, The region on chromosome 16 (chr16: 26.7-27.0 Mb) was zoomed in, sequentially illustrating the corresponding GWAS results, F_{ST} , π , and the situation of the corresponding genes. E, Linkage disequilibrium analysis was conducted for the gene *Vitvi031045* and an upstream segment (chr16:26.92-26.93 Mb), with the upper section revealing the P values of the GWAS results.

Integration of GWAS and RNA-seq identified pest-resistant candidate genes

In order to verify the accuracy of the DL model predictions and GWAS analysis results, we conducted transcriptomic analyses. Under assault by the two-spotted spider mite, *Tetranychus urticae*, another herbivorous insect, we identified a total of 3979 differentially expressed genes (DEGs) with $P \leq 0.05$ and a fold change greater than 2. Subsequently, a total of 2010 upregulated genes and 1969 downregulated genes were identified (Fig. 4A and Supplementary Table 10). GO enrichment analysis revealed these genes are significantly involved in signal transduction, defense responses, and metabolic processes, consistent with previous studies³² (Supplementary Table 11). It is noteworthy that DEGs were also found to be enriched in a cluster with the highest significance, specifically in protein phosphorylation and protein kinase activity (protein serine/threonine/tyrosine kinase activity) ($P < 0.05$), in line with our GWAS analyses. We observed a convergence between the DEGs uncovered through RNA-seq and the genes associated with binary traits (BTGs) and continuous traits (CTGs). There were 13 genes common to both traits, with 7 genes shared between BTGs and DEGs, and 12 genes shared between CTGs and DEGs, respectively (Fig. 4B). Notably, all three analyses identified the presence of two genes (*Vitvi034587* and *Vitvi031046*). *Vitvi031046* (*SAUR32*) encode an auxin-responsive protein and *Vitvi034587* (*ERF110*), known as *ethylene-responsive transcription factor* (*ERF110*, *Vitvi034587*), which was significantly induced under conditions of herbivore attack (Fig. 4C). *Vitvi012987* (*ATHB-6*) and *Vitvi031015* (*DMR6*)

constituted the intersection of BTGs and DEGs, playing roles in the regulation of the ABA signaling pathway and the synthesis of defensive oxygenase, respectively^{33,34}. They were also significantly induced in expression, indicating their relevance to pest defense (Fig. 4C). Within the intersection of CTGs and DEGs, two key genes have been identified (*Vitvi021951* and *Vitvi021953*). These genes are associated with continuous traits (snp_12__9025204) in GWAS and are strongly induced under herbivore attack. Additionally, it is noteworthy that these three genes are tandemly linked (Chr12: 9.01M-9.04M) and are encoded by the same gene, *calcium-transporting ATPase 12, ACA12* (Fig. 4D). *ACA12* encodes a calcium ion pump localized on the plasma membrane of plant cells (Fig. 4E). Mutants of *ACA12* have been demonstrated to exhibit extreme sensitivity to herbivorous attack by *Spodoptera littoralis*. This sensitivity manifests as a significant increase in leaf damage, accompanied by a decrease in intracellular calcium ion concentrations³⁵. Simultaneously, *Vitvi031045* (previously mentioned in the GWAS results) was also found to be significantly induced (Fig. 4C, E). Analysis of integrated transcriptome data demonstrated the effectiveness of the identified pest resistance genes across both traits through the application of DL models.

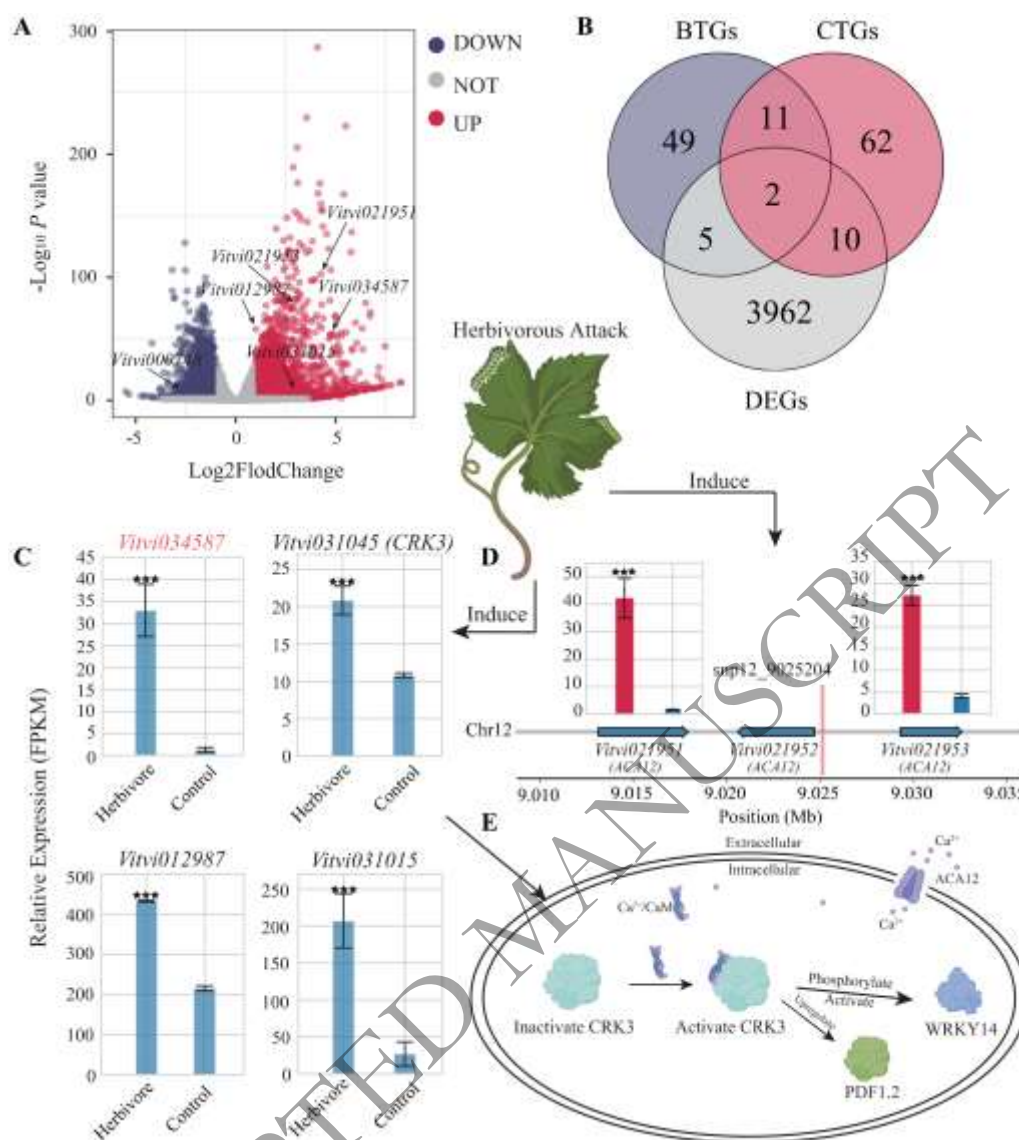


Fig. 4 | Transcriptomic analyses in grapevine under herbivorous and normal conditions. **A**, Volcano plot of differential gene expression, with red indicating upregulation and blue indicating downregulation, the genes indicated by the arrows illustrate that they are also located by the GWAS results. **B**, Venn diagram illustrating the overlap between genes associated with binary and continuous phenotypes, as well as differentially expressed genes. **C**, Gene expression patterns of the four pest-defense genes which derived from the intersection of GWAS and RNA-seq results. **D**, Three genes in the region of chromosome 12 (9.010M-9.035 Mb) encode the same protein and were detected by GWAS (snp12_9025204). Bar charts for each gene depict the corresponding expression levels, where red indicates herbivory conditions, and asterisk denote significant differences ($\star$: $P < 0.05$, $\star\star\star$: $P < 0.001$). **E**, The functions performed by two insect-resistant genes within the cell are elucidated, along with their potential interactions.

369 **Improving genomic selection with the optimal machine learning model**

370 The outcomes from GWAS with binary and continuous traits were utilized to
 371 construct genomic selection models based on machine learning methodologies,
 372 respectively. Among four evaluated models, including Logistic Regression, Support
 373 Vector Classifier (SVC), Random Forest Classifier (RFC) and Naïve Bayes Classifier
 374 models for the binary trait prediction (Supplementary Table 12). The performance of
 375 the RFC model with different variations consistently fell short, achieving an accuracy
 376 that did not exceed 85%. Naive Bayes outperformed RFC but fell behind the other
 377 two models, achieving a maximum average accuracy of 87.6%. As the number of
 378 variations increased, the performance of the SVC model consistently improved.
 379 Beyond 500 variations, the average accuracy showed a gradual plateauing, reaching
 380 its peak at 93% when the number of variations reached 5000. The logistic regression
 381 model was deemed more suitable for this modeling task. In contrast to the SVC model,
 382 logistic regression reached its peak accuracy at 2000 variations (94.5%) and exhibited
 383 a declining trend in accuracy with 5000 variations (Fig. 5A). We performed the final
 384 validation of the test set by combining the logistic regression model with the
 385 appropriate number of variations. The accuracy achieved was 95.7%, and a significant
 386 correlation of 0.94 ($P = 1.4e-11$) was observed between the predicted values and the
 387 actual phenotypes (Fig. 5B). In the prediction of continuous traits, we compared
 388 various linear models (Lasso, Ridge, ElasticNet) and machine learning models,
 389 including Random Forest Regression (RFR) and Support Vector Regression (SVR)
 390 (Supplementary Table 13). Among these models, the Lasso model exhibited a poor
 391 performance, with the highest average correlation coefficient not exceeding 0.72. The
 392 accuracy of the RFR showed negligible variation, reaching a maximum of only 0.79.
 393 ElasticNet performed well, achieving a high accuracy of 0.88, and demonstrated
 394 stable performance beyond 500 variations. According to the results, the Ridge model
 395 exhibited instability, with a sudden increase in accuracy after finding an appropriate
 396 number of variations, reaching a peak of 0.89 at 10,000 variations. SVR outperformed

all other models, showing a steady increase in performance. Beyond 2000 variations, its performance gradually stabilized, reaching a peak accuracy of 0.92. (Fig. 5C). Based on the results of CV, we selected SVR and applied the identified 10,000 SNPs to predict the continuous phenotype in the test set. Ultimately, we obtained a significant correlation coefficient of 0.90 ($P = 3.5e-9$) between the predicted and actual phenotypes, demonstrating the accuracy of the predictions (Fig. 5D). Additionally, all Support Vector Machine models (SVC, SVR) undergo CV to select the optimal kernel function that represents the corresponding optimal results (Supplementary Fig. 10).

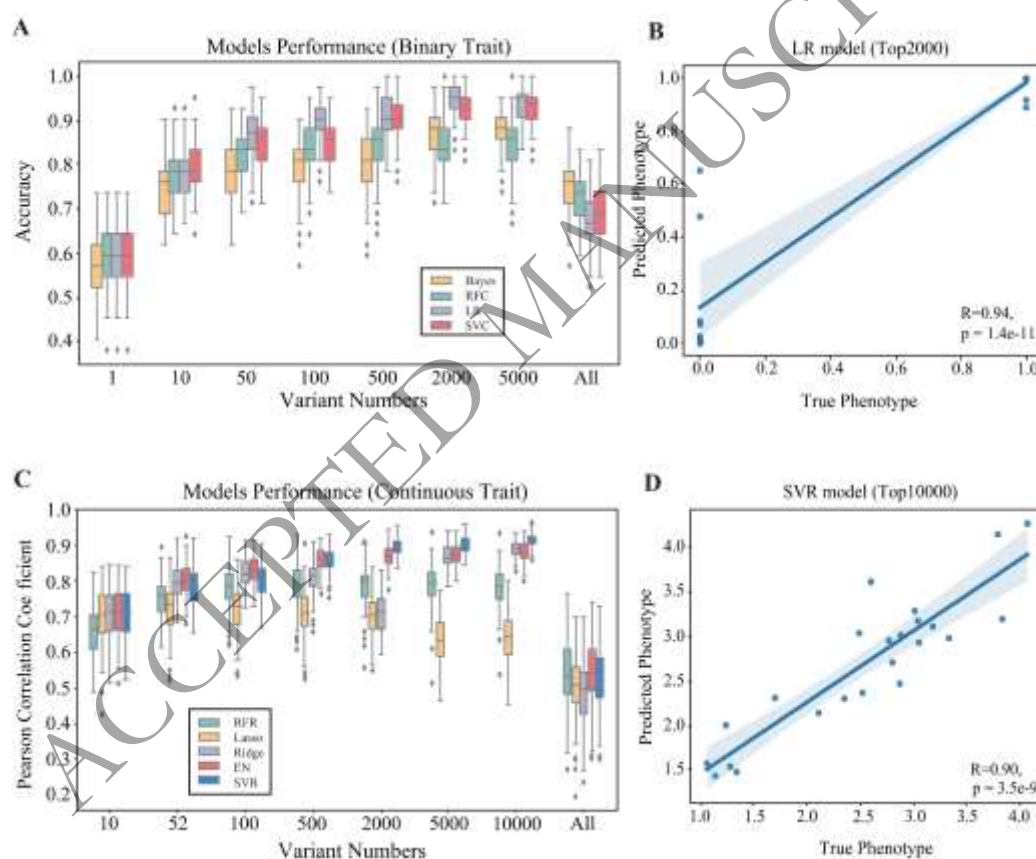


Fig. 5 | Machine learning based genomic selection of pest-resistance in grapevine. **A**, Cross validation of the accuracy of the four classification models, using the average accuracy value as the accuracy indicator. **B**, Using the optimized model and variable count, predictions were made on the test set samples, revealing the correlation between the predicted binary phenotype and the true binary phenotype. **C**, Cross validation of the accuracy of the five regression models, using the average correlation coefficient value as the accuracy indicator. **D**, Using the

optimized model and variable count, predictions were made on the test set samples, revealing the correlation between the predicted continuous phenotype and the true continuous phenotype.

Discussion

Our study implemented cutting-edge breeding methodologies to develop grapevine cultivars with enhanced resistance to pests, leveraging recent breakthroughs in DL based plant phenomics and ML-based genomic selection. Specifically, we employed GWAS to map the genetic basis of pest resistance in grapes^{36,37}. Subsequently, genomic selection is utilized to selectively breed individuals harboring pest-resistant alleles. Although binary phenotypes are commonly deployed in pest resistance analyses, they overlook subtle-effect genes in GWAS. Conversely, the continuous phenotype generated by DCCN-PDS enables the detection of minor-effect genes, thereby offering a more comprehensive understanding of pest resistance mechanisms. We integrated DL based binary and continuous pest damage traits, leading to the identification of 139 candidate genes endowing pest resistance, facilitating machine learning based genomic prediction of pest damage traits in grapevine breeding.

The genetic basis of pest-resistance in grapevine

This study aimed to identify genes associated with resistance to *Spodoptera litura* *Fabricius* (SLF) in grape. Since there was no prior knowledge about the impact of identified genes on SLF resistance, both binary and continuous damage scores were used as phenotypes in GWAS. Our findings revealed 139 candidate genes related to insect resistance that participate in various signaling pathways, including jasmonic acid (JA), salicylic acid (SA), ethylene, and others. Notably, the gene *CRK3* showed a significant association with both types of phenotypes, indicating its crucial role in SLF resistance. Previous research has also suggested *CRK3*'s involvement in plant defense responses, as overexpressing *CRK3* in *Arabidopsis* enhances defense properties, while knockout mutants exhibited reduced expression levels of the defense gene *PDF1.2*³¹. Additionally, *CRK3* has been demonstrated to phosphorylate tyrosine residues of herbivory-responsive regulators like *WRKY14*, highlighting its

participation in the defense response against SLF³⁸. It is worth mentioning that *ACA12*, another gene associated with continuous damage scores, encodes a calcium ion pump protein localized to the plasma membrane of plant cells³⁹. Research has shown that *ACA12* mutants are more susceptible to SLF attacks compared to wild-type plants, suggesting that *ACA12* may play a vital role in providing a calcium ion-enriched environment for *CRK3*³⁵. This interaction could potentially activate *CRK3* or other defense proteins reliant on calcium ion signaling, initiating subsequent defense responses⁴⁰. Nonetheless, further investigation is necessary to validate this hypothesis. It is interesting to note that genes conferring resistance to diseases tend to display higher levels of polymorphism compared to other genomic regions, likely due to the selective pressure imposed by rapidly evolving pathogen populations³⁹. Unfortunately, only a minuscule proportion (approximately 0.1%) of the available biodiversity in resistance loci controlling pests and pathogens has been leveraged in commercial crop varieties. To develop more pest-resistant crops, integrating diverse genetic resources from local varieties, wild ancestors, and wild relatives into breeding programs is essential⁴¹. These resources possess distinct combinations of resistance traits shaped by evolutionary pressures, rendering them valuable sources of novel genetic material⁴². By capitalizing on the innate resistance of host plants through these resources, breed selection becomes more effective, and germplasm utilization is significantly improved, ultimately facilitating crop improvement.

DL/ML based methods for grapevine genomic breeding

Compared to traditional methods, deep learning enables more efficient, objective and accurate assessment of phenotypic variation in plants^{43,44}. Commonly, the extent of leaf damage is manually assessed by visual inspection therefore it is slow, labor-consuming and subjective to the judgement of individual assessors. Our DL models can objectively assess one leaf image within 0.03 to 0.04 seconds and measure its damage condition with high accuracy, up to 0.95 for binary phenotype and with a high correlation, up to 0.94 for continuous scores, compared to the manual labeling. Therefore, the sample size can be easily scaled up if required. Our DL models

improve the feasibility to establish large scale phenomics and empower the molecular breeding of grapes⁴⁵. Noteworthy, in insect pest research, DL is mostly applied to identify the pest types based on the morphological features of the pest itself⁴⁶, hence not designed to evaluate the pest resistance of the plants. To precisely determine the resistance feature of plants themselves, it is required to measure the damage conditions in various infestation parts of the different plant samples, for example damage to air potato tubers⁴⁷, leaf damage⁴⁸, dead heart tillers and whitehead-panicles⁴⁹, roots damage⁵⁰ and so on. In our study, the major pest in grape cultivation is SLF. Its larvae mainly consume grape leaves and can cause complete defoliation⁵¹. Generally, the damage to plants caused by the disease and pest are often considered together in most studies using popular DL models designed for image recognition⁵². It is difficult to disengage the disease and pest resistant capabilities from the same plant. To avoid this, our study is deliberately based on disease-free plants and our model is then capable of measuring the damage (mild vs. severe) caused by the pest only. We selected VGG16, which outperformed the other widely used image recognition models, including AlexNet, ResNet50, ResNet101, InceptionV3, and DenseNet121, as our DL models backbone. We added four residual networks based on the backbone to generate both binary and continuous damage scores. For genomic selection (GS) of pest-resistant grapevines, we compared multiple ML methods including random forest, naive Bayes, support vector machines, elastic net and other models, to avoid empirical bias in model selection. We finally selected logistic regression (95.7 % in accuracy) and support vector regression (0.90 in correlation coefficient) as models for precisely predicting binary and continuous phenotypes, respectively.

Pest resistance is a polygenic complex traits, with many small-effect variants contribute jointly to phenotypic variation. Moreover, the predictive ability of significant SNPs may be limited by population-specific differences and genomic landscapes (e.g. genomic linkage, 3D structure and epigenomic modifications). In summary, filtering markers through the GWAS approach can enhance predictive performance. Future studies could further explore how combining functional annotations or biological prior knowledge could optimize marker selection to improve

prediction stability and transferability. Furthermore, although this study demonstrates the effectiveness of machine learning methods in predicting pest resistance traits, we acknowledge that traditional genome-wide selection methods, such as GBLUP, remain widely adopted benchmark methods in plant breeding. As highlighted in the review by Tong and Nikoloski, GBLUP and related mixed linear models have shown predictive ability for various agronomic traits across multiple crop species⁵³. The focus on alternative machine learning methods is due to their ability to capture non-additive genetic effects and complex genotype-phenotype relationships, which may be particularly important for pest resistance traits. Future research could involve direct comparisons with GBLUP and more recent deep learning-based genome selection methods, such as DNNGP, to comprehensively assess the relative advantages of different modeling frameworks.

In conclusion, our results indicate that deep learning is an effective technique for phenotypic evaluation with high speed and low biases. It lays the foundation for large-scale phenomics data establishment. Our integrative analysis of genomics, phenomics, transcriptomics combined with the well selected ML model with high prediction accuracy, our study explored the polygenic basis of pest-resistance, facilitating efficient breeding for pest-resistance grapevines. The phenomics evaluate model and genomic selection model would be integrated to automated breeding platform. The breeding strategy will be greatly advanced by the integration of mass breeding data and Artificial intelligence (AI) algorithms.

Materials and Methods

Viticulture conditions and pest infestations outbreak

A total of 231 grape varieties were introduced from the Grape Germplasm Repository of the Zhengzhou Fruit Research Institute of the Chinese Academy of Agricultural Sciences and planted in the Grape Germplasm Resource Nursery of the Shenzhen Institute of Agricultural Genomics (AGIS) of the Chinese Academy of Agricultural Sciences (CAAS) in Nov. 2021. All experimental varieties were 2-year-old cuttings

from standard vines and were cultivated in pots under greenhouse conditions, with each pot including at least three branches to represent given accession. All plants were subject to standard management practices prior to the experiment, including cultivation, irrigation, fertilization, pruning, and disease control (Supplementary Fig. 6). To examine the impact of different grape varieties on pest damage, we randomly arranged different grape varieties in a greenhouse. We performed a ~20 days experiment, when a natural outbreak of tobacco cutworm occurred in our greenhouse during July-August 2022, with 9-10 larvae per plant. The pesticide was sprayed after the experiment ended (Fig. 2). We randomly take pictures for six to ten leaves for each accession in early September 2022 to represent the pest damage of the sample.

Collection and preprocessing of image datasets for model training

The images were captured with the high-quality built-in camera of a Sony ZV-1 II, which features advanced sensor technology and image processing. We captured thousands of grape leaf samples from different locations, including wild and farm environments (<https://github.com/zhouyflab/Pest-Resistance>), showcasing a range of colors, backgrounds, and pest damage levels. To focus on the main part of the leaves and eliminate excess pixels, we trained an a YOLO object detection model for leaf detection⁵⁴. This allowed us to accurately and automatically crop leaf bounding boxes from thousands of images (Fig. 1c). Although the model has the capability to detect wormholes as well, they were not considered in this study, and the weights file for this YOLO model can be available for download (<https://github.com/zhouyflab/Pest-Resistance>). The object detection model achieved an impressive mAP of 0.91 after 300 epochs of training (Supplementary Fig. 11). Leave images were resized to 224x224 pixels using the Python Imaging Library (PIL) (Version:9.4.0), and pixel values were normalized to the 0-1 range for consistent process and model input. To increase data diversity and generalization, we used the ImageDataGenerator class from Keras (Version:2.4.0) to apply random transformations and augmentations, including random rotation up to 45 degrees, shifts of 15% in width and height, horizontal flipping, and random zooming with a range of 50%⁵⁵.

Image labeling and dataset splitting process

To train the binary classification model, a total of 1810 images were labeled using one-hot encoding: leaves with over 25% pest damage were labeled as 1, while those with 25% or less damage were labeled as 0. The dataset was randomly split into a 4:1 ratio, with 80% of the data (1448 images) used for CV (training and validation), and the remaining 20% (362 images) used as the test set. For the regression model, a total of 2620 images were used: 439 images for testing, while the remaining 2181 images for training and CV process, split in a 4:1 ratio. All the images are labeled from 1 to 5 based on damage severity: 1 for up to 20% damage area, increasing by 1 for each additional 20%, with 5 for damage exceeding 80%. Note that all labels, including binary classification labels and regression labels, are determined among three raters to reduce subjective bias (Supplementary Fig. 12).

Modeling process of six binary task models

All deep learning model for grape leaf recognition were built using the Keras library with TensorFlow (Version: 2.4.0) backend⁵⁶. Modeling was performed on a Linux platform with a GeForce RTX 3090 GPU, equipped with 24 GB of memory.

For the binary-classification modeling procedures, a comparative analysis of the performance of the AlexNet⁵⁷, VGG16⁵⁸, ResNet50, ResNet101⁵⁹, InceptionV3⁶⁰, and DenseNet121⁶¹ models was conducted for the classification of pest damage into two categories: mild (0) and severe (1). The network architecture was modified with a focus on the fully connected and output layer. Each model featured two fully connected layers with 4096 neurons and relu activation, and a dropout rate of 0.5 was applied to enhance robustness⁶². The output layer used a sigmoid activation function to match the binary cross-entropy loss. To prevent overfitting, early stopping was implemented to terminate the training process when it no longer resulted in meaningful improvements. Models varied in complexity, resulting in different patience values. For example, AlexNet was tolerated for 20 epochs, while DenseNet121 required 40 epochs. We evaluated these models using 5-fold CV and

calculated accuracy and F1-score metrics with scikit-learn (Version 1.0.2). Random seeds were applied to ensure consistent data splitting during the CV process across all models. All the six models share the same hyperparameters range and settings. We only tuned the batch size (32, 64, 128) and learning rate (0.0001, 0.001, 0.01), while the number of epochs (200) was not specifically adjusted due to early stopping. Then Adam optimizer was used for all models^{63,64}. Transfer learning is applied to enhance model performance⁶⁵, with the weights automatically downloaded through the Keras library.

DCNN-PDS modeling procedures

The regression model generates a continuous value to assess the pest-damage score (PDS) of grape leaves. The model was built using the Keras library⁶⁶ (Version:2.4.0) and was executed on a Linux platform. The network architecture is constructed based on a combination of VGG16⁵⁸ and residual networks⁵⁹. VGG16 is renowned for its strong feature extraction capabilities. However, merely stacking networks to boost performance can lead to the loss of crucial image information and degrade the overall model⁵⁹. To address this, we integrated residual networks. By adding four residual blocks after the VGG16 convolutional layer, we could stack deeper CNNs without worrying about network degradation. Each residual block contains two convolutional layers with 512, 256, 128, and 64 filters, respectively. A global average pooling (GAP) layer was then added after the last convolutional layer to reduce feature dimensions. Three fully connected (FC) layers with ReLU activations were appended behind the GAP layer for feature extraction and fusion, each with 1024 neurons. A dropout layer with a 0.3 rate followed each FC layer to reduce overfitting. Unlike the original VGG16, which used softmax for multi-classification, our model aims to solve a numerical regression problem for pest-damage scores⁵⁸. Therefore, we used sigmoid activation in the final FC regression layer⁶⁷, with the output X5 serving as the pest damage score, corresponding to the manual label value (Fig. 1c).

In this study, the parameters of the convolutional layers of DCNN-PDS have been pre-trained on ImageNet⁶⁸ and are frozen during the training process⁶⁹. For this

regression model, mean squared error (MSE) was chosen as the loss function, and mean absolute error (MAE) was used as the performance metric to evaluate accuracy. Similarly, we fine-tuned the learning rate (0.0001,0.001,0.01) and batch size (32,64,128) through 5-fold CV with early stopping set to a tolerance of 35 epochs. The overall training process consisted of 300 epochs, after CV, we proceeded with formal model training and evaluated it on a new test set, keeping the optimal hyperparameters unchanged. To further assess the performance of the model, we utilized the Numpy library (Version: 1.19.5) to calculate the Pearson correlation coefficient⁷⁰ between the predicted values and the labels, aiming to determine their linear relationship.

Grape phenotype calculation

Six to ten images are available for each grape accession to determine its phenotype (<https://github.com/zhouyflab/Pest-Resistance>). To calculate the binary phenotype, indicating the severity or mild of pest damage, the model computes a probability value of the 0-1 classification for each image. Images with an average probability value below 0.5 are classified as mild pest damage, while those with an average probability value greater than 0.5 are classified as severe pest damage. For a more detailed evaluation of pest damage, the DCCN-PDS provides direct damage scores for multiple images. These scores are then averaged to obtain the final continuity phenotype value. The accurate scores provided by DCCN-PDS quantify the severity of pest damage on a continuous scale from 0-5, without dimensions. This approach offers a more objective alternative to manual measurement and provides data suitable for genetic analysis.

Distribution statistics of pest phenotypes

The Sapiro-Wilk test was employed to assess the normality of the sample data. Specifically, the shapiro.test was utilized in R to compute the W-statistic and associated p-value for various samples. A p-value inferior to 0.05 signifies that the sample's phenotypic distribution conforms to a normal distribution. In order to

evaluate variance homogeneity on normally distributed data, the Bartlett test was performed. The bartlett. The test function in R was leveraged to determine the corresponding p-value for each sample. If the p-value surpasses 0.05, it may be inferred that the variance of the data exhibits homogeneity. We employed R's Spearman and rcorr functions to calculate Spearman's footrule distance and correlation, respectively. Spearman's footrule is a non-parametric statistical method used to quantify the difference between two ranked sequences. It is defined as the sum of the absolute differences in positions for each element in the two lists, and its values range from 0 to $n^2/2$ (where n represents the sample size). List A was based on the order of the phylogenetic tree, while List B was based on the degree of pest damage to different varieties. Furthermore, we generated 100 random orders for conducting a single-sample t-test.

The GWAS analyses

In this study, paired-end resequencing reads were mapped to the *V. vinifera* reference genome PNT2T⁷¹ using BWA-mem2 (Version: 2.2.1) with default parameters. Mapping results were converted into the BAM format and filtered for unmapped and non-unique reads using SAMtools (Version: 1.17). BWA alignment was conducted, and two functions "vc and joint" of GTX (Version: 2.2.1) were applied to obtain SNP VCF files. To ensure the accuracy of the results, SNPs with a missing genotype frequency greater than 0.05 or a minor allele frequency (MAF) less than 0.05 were excluded from analysis. Imputation was not performed. Populations were considered for their structure and cryptic relationships, and GEMMA⁷² (Version: 0.98.3) was used to implement the GWAS of the standard linear mixed model:

$$y = W\alpha + X\beta + u + \varepsilon.$$

The whole-genome significance cutoff was defined using the adjusted Bonferroni test threshold, which was set as $P < 1e-5$.

Transcriptomic analyses under herbivorous conditions

Our transcriptome data sets are originated from the Sequence Read Archive (SRA) accession SRP067967, which was sequenced for its transcriptome data on grape leaves under herbivorous exposure to two-spotted spider mite (*Tetranychus urticae*)³². The raw transcriptomic data was processed using fastq-dump⁷³ (Version:2.11.0), converting it into paired-end fastq files. Quality control was performed using FastQC⁷⁴ (Version:0.11.9) to assess data quality. Subsequently, trim-galore⁷⁵ software (Version:0.6.7) was utilized to filter out adapter sequences and low-quality reads (--Stringency 3 --phred33 -q 25). The data was then aligned to the grape reference genome using STAR⁷⁶ software (Version:2.7.10b), resulting in aligned BAM files. Quantification of gene expression was performed using featureCounts (Version:2.0.1). For the final analysis, we employed DESeq2⁷⁷ package (Version:1.38.3) in RStudio to conduct differential gene expression analysis. Genes with a fold change greater than two and a *P* value less than 0.05 will be considered as differentially expressed genes (DEGs). All GO enrichment analyses were conducted on the DAVID website (<https://david.ncifcrf.gov/tools.jsp>). Additionally, we employed materials from BioRender (<https://app.biorender.com/>) to construct mode pattern.

The comparison of genomic selection models based on the Cross-validation

Regarding the severity of pest damage, two phenotypic descriptors were estimated. One was binary qualitative shape with the phenotype divided into two classes (“mild” vs. “severe”). The other one represented a quantitative measurement of the damage with the phenotype quantified into a continuous scale ranging from 0 to 5. For binary phenotype, Logistic Regression, Support Vector Classifier (SVC), Random Forest Classifier (RFC) and Naive Bayes Classifier models were selected to enable comparisons of predictive accuracy. The prediction accuracy between different models was determined by the average accuracy. For continuous phenotype prediction, two machine learning models, Support Vector Regressor (SVR), Random Forest Regressor (RFR), and three regularization models, Lasso (L1 regularization), Ridge regression (L2 regularization) and ElasticNet (combining L1 and L2 regularization), were compared between each other. L1 and L2 regularization represent two types of

marker effect estimation. The former, L1 regularization, is well-suited for identifying and adapting to a few major QTLs that have significant effects on the phenotype, L2 regularization, also known as ridge regression, is effective in capturing the effects of many minor-effect QTLs. The Pearson's correlation coefficient between the predicted and the observed phenotypes were calculated as the prediction accuracy. The support vector machine (SVM) models were applied to predict both binary and continuous phenotypes. Additionally, SVM offers various kernel functions, including the linear kernel, polynomial kernel, and radial basis function kernel. These three kernel functions were also compared by the CV to select the optimal one that represented the best results for this model. The all models utilized in genomic selection were implemented using the scikit-learn Python library (Version: 1.0.2). Different models in two types of pest phenotypes were compared based on accuracy and the optimal model was chosen.

Stepwise variation selection based on GWAS results for optimal prediction

Regarding the results from GWAS, the sorting based on their *P* values and a stepwise selection of variations were performed to identify the optimal number of variations to construct a genomic selection model. Before modeling, Plink (Version:1.90b7) was used to filter variations with the parameters including window size set to 100, step size set to 50, and linkage disequilibrium (LD) threshold set to 0.2 to reduce redundancy and minimize the impact of LD in the genomic selection process. After filtering, a total of 7,359,942 variations were reduced to 313,961 variations, representing all the variations participated in the modeling. Different numbers of variations in models with binary and continuous phenotypes were evaluated with cross validation. With the binary phenotype, a total of 50 significant SNPs ($P < 1e-5$) were identified through GWAS. The variations with the numbers including 1, 10, 50 (all significant SNPs), 100, 500, 2000, 5000, and 313,961 (all variations) according to their sorted *P* values were selected for genomic selection. Regarding the continuous phenotypes, a total of 52 significant SNPs ($P < 1e-5$) were identified. The numbers of variations including 10, 52 (all significant SNPs), 100, 500, 2000, 5000, 10000, and

313,961 (all variations) were considered for prediction modeling. 10% of the samples' phenotype data were reserved for testing, while the remaining 90% of the samples were split in a 4:1 ratio for modeling and validation. The optimal number of variations was evaluated based on the final modelling accuracy averaged over one hundred iterations. Random seeds were set to ensure the training data consistency across these 100 modeling and evaluation iterations.

Data availability

All resequencing data generated have been deposited in the NCBI database under the BioProject: PRJNA994294.

Code availability

All scripts performed in this study are available on Github:

<https://github.com/zhouyflab/Pest-Resistance>

Acknowledgements

We thank members of the Zhou lab at AGIS for discussion and comments on the project.

Author contributions

Y.Z. designed the project. Y.Z., Z.L. and Y.W. supervised the project. Y.G., Z.L. design and train image recognition and object detection models. Y.G., F.Z., and Q.X. performed the bioinformatic analyses. X.W., H.X., W.L., L.S., Y.Z., and C.L. assisted in bioinformatics analyses., Y.G., X.S., and W.M. provided essential technology support. Q.L. and A.M. provided the plant material for genome analyses. Y.G., Z.L., and Y.W. wrote the draft. Y.L., X.F., X.F., C.Z., P.Z., C.L., and Z.Z. revised the manuscript. All authors contributed to manuscript preparation and read, commented on, and approved the manuscript.

Competing interests

The authors declare no competing interests.

Funding

This work was supported by the National Key Research and Development Program of China (No. 2023YFD2200702), the project of National Key Laboratory for Tropical Crop Breeding (No. NKLTCB202325), the National Natural Science Foundation of China (No. 32372662) and the Science Fund Program for Distinguished Young Scholars of the National Natural Science Foundation of China (Overseas) to Yongfeng Zhou.

References

1. E. C. Oerke. Crop losses to pests. *The Journal of Agricultural Science* **144**, 31-43 (2006).
2. J.-C. Simon&J. Peccoud. Rapid evolution of aphid pests in agricultural environments. *Current opinion in insect science* **26**, 17-24 (2018).
3. Z. He, J. F. Guo, S. R. Reitz, Z. R. Lei&S. Y. Wu. A global invasion by the thrip, *Frankliniella occidentalis*: Current virus vector status and its management. *Insect science* **27**, 626-645 (2020).
4. J. Beddington. Food security: contributions from science to a new and greener revolution. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365**, 61-71 (2010).
5. I. Mahmood, S. R. Imadi, K. Shazadi, A. Gul&K. R. Hakeem. Effects of pesticides on environment. *Plant, soil and microbes: volume 1: implications in crop science* 253-269 (2016).
6. A. J. Forgash. History, evolution, and consequences of insecticide resistance. *Pesticide biochemistry and physiology* **22**, 178-186 (1984).
7. J. A. Gatehouse. Plant resistance towards insect herbivores: a dynamic interaction. *New phytologist* **156**, 145-169 (2002).
8. A. R. War, M. G. Paulraj, M. Y. War&S. Ignacimuthu. Herbivore-and elicitor-induced resistance in groundnut to Asian armyworm, *Spodoptera litura* (Fab.)(Lepidoptera: Noctuidae). *Plant Signaling & Behavior* **6**, 1769-1777 (2011).
9. A. Kundu, S. Mishra&J. Vadassery. *Spodoptera litura*-mediated chemical defense is differentially modulated in older and younger systemic leaves of *Solanum lycopersicum*. *Planta* **248**, 981-997 (2018).
10. A. R. War.et al. Mechanisms of plant defense against insect herbivores. *Plant signaling & behavior* **7**, 1306-1320 (2012).
11. A. Verhage, S. C. van Wees&C. M. Pieterse. Plant immunity: it's the hormones talking, but what do they say? *Plant physiology* **154**, 536-540 (2010).
12. M. Ahmad, A. Ghaffar&M. Rafiq. Host plants of leaf worm, *Spodoptera litura* (Fabricius)(Lepidoptera: Noctuidae) in Pakistan. *Asian J. Agric. Biol* **1**, 23-28 (2013).

13. E. P. o. P. Health.et al. Pest categorisation of *Spodoptera litura*. *EFSA Journal* **17**, e05765 (2019).
14. R. Balikai&Y. Kotikal. Pest status of grapevine in northern Karnataka. *Agricultural Science Digest* **23**, 276-278 (2003).
15. S. Jagginavar, A. Biradar&A. Bagali. Menace of *Spodoptera litura* Fab. on grape. *Insect Environment* **5**, 76-77 (1999).
16. X. Wang.et al. Integrative genomics reveals the polygenic basis of seedlessness in grapevine. *Current Biology* **34**, 3763-3777. e5 (2024).
17. Z. Liu.et al. Grapevine pangenome facilitates trait genetics and genomic breeding. *Nature Genetics* 1-11 (2024).
18. H. Xiao.et al. Impacts of reproductive systems on grapevine genome and breeding. *Nature Communications* **16**, 2031 (2025).
19. C. H. Bock, K.-S. Chiang&E. M. Del Ponte. Plant disease severity estimated visually: a century of research, best practices, and opportunities for improving methods and practices to maximize accuracy. *Tropical Plant Pathology* **47**, 25-42 (2022).
20. M. Laxa, M. Liebthal, W. Telman, K. Chibani&K.-J. Dietz. The role of the plant antioxidant system in drought tolerance. *Antioxidants* **8**, 94 (2019).
21. C. R. S. Buffara.et al. Elaboration and validation of a diagrammatic scale to assess downy mildew severity in grapevine. *Ciência Rural* **44**, 1384-1391 (2014).
22. R. K. Varshney.et al. Fast-forward breeding for a food-secure world. *Trends in Genetics* **37**, 1124-1136 (2021).
23. Y. LeCun, Y. Bengio&G. Hinton. Deep learning. *nature* **521**, 436-444 (2015).
24. S. P. Mohanty, D. P. Hughes&M. Salathé. Using deep learning for image-based plant disease detection. *Frontiers in plant science* **7**, 1419 (2016).
25. S. Sladojevic, M. Arsenovic, A. Anderla, D. Culibrk&D. Stefanovic. Deep neural networks based recognition of plant diseases by leaf image classification. *Computational intelligence and neuroscience* **2016**, (2016).
26. A. Samanta, G. Das&S. K. Das. Roles of flavonoids in plants. *Carbon* **100**, 12-35 (2011).
27. J. R. Ecker&R. W. Davis. Plant defense genes are regulated by ethylene. *Proceedings of the National Academy of Sciences* **84**, 5202-5206 (1987).
28. P. L. Rodriguez. Protein phosphatase 2C (PP2C) function in higher plants. *Plant molecular biology* **38**, 919-927 (1998).
29. A. Schweighofer.et al. The PP2C-type phosphatase AP2C1, which negatively regulates MPK4 and MPK6, modulates innate immunity, jasmonic acid, and ethylene levels in Arabidopsis. *The Plant Cell* **19**, 2213-2224 (2007).
30. J. C. Jeong.et al. Isolation and characterization of a novel calcium/calmodulin-dependent protein kinase, AtCK, from arabidopsis. *Molecules & Cells (Springer Science & Business Media BV)* **24**, (2007).
31. T. Miyamoto.et al. Tyrosine Kinase-Dependent Defense Responses Against Herbivory in Arabidopsis. *Frontiers in Plant Science* **10**, (2019).
32. J. Díaz-Riquelme.et al. Comparative genome-wide transcriptome analysis of *Vitis vinifera* responses to adapted and non-adapted strains of two-spotted spider mite, *Tetranychus urticae*. *BMC genomics* **17**, 1-15 (2016).
33. M. Van Damme, R. P. Huibers, J. Elberse&G. Van den Ackerveken. Arabidopsis DMR6 encodes a putative 2OG-Fe (II) oxygenase that is defense-associated but required for susceptibility to downy

- mildew. *The Plant Journal* **54**, 785-793 (2008).
34. A. Rajavel.et al. Unravelling the complex interplay of transcription factors orchestrating seed oil content in *Brassica napus* L. *International journal of molecular sciences* **22**, 1033 (2021).
35. N. Fotouhi.et al. ACA pumps maintain leaf excitability during herbivore onslaught. *Current Biology* **32**, 2517-2528. e6 (2022).
36. A. Karn.et al. Discovery of the REN11 locus from *Vitis aestivalis* for stable resistance to grapevine powdery mildew in a family segregating for several unstable and tissue-specific quantitative resistance loci. *Frontiers in Plant Science* **12**, 733899 (2021).
37. T. Possamai&S. Wiedemann-Merdinoglu. Phenotyping for QTL identification: A case study of resistance to *Plasmopara viticola* and *Erysiphe necator* in grapevine. *Frontiers in Plant Science* **13**, 930954 (2022).
38. Q. Zhu.et al. The Arabidopsis AP2/ERF transcription factor RAP2. 6 participates in ABA, salt and osmotic stress responses. *Gene* **457**, 1-12 (2010).
39. M. Klimecka&G. Muszyńska. Structure and functions of plant calcium-dependent protein kinases. *Acta Biochimica Polonica* **54**, 219-233 (2007).
40. T. Miyamoto.et al. Tyrosine kinase-dependent defense responses against herbivory in Arabidopsis. *Frontiers in plant science* **10**, 776 (2019).
41. A. Goverse, G. Smant, L. Bouwman, E. Bakker&J. Bakker. The BIOEXPLOIT project. *Potato research* **52**, 209-213 (2009).
42. H. Xiao.et al. Adaptive and maladaptive introgression in grapevine domestication. *Proceedings of the National Academy of Sciences* **120**, e2222041120 (2023).
43. M. P. Pound.et al. Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. *Gigascience* **6**, gix083 (2017).
44. S. Arya, K. S. Sandhu, J. Singh&S. Kumar. Deep learning: As the new frontier in high-throughput plant phenotyping. *Euphytica* **218**, 47 (2022).
45. Z. Tang.et al. Affordable High Throughput Field Detection of Wheat Stripe Rust Using Deep Learning with Semi-Automated Image Labeling. *Computers and Electronics in Agriculture* **207**, 107709 (2023).
46. H. Guan.et al. A lightweight model for efficient identification of plant diseases and pests based on deep learning. *Frontiers in Plant Science* **14**, (2023).
47. F. A. Dray Jr&S. L. Goldstein. A novel method for quantifying insect damage to plant storage organs. *Entomologia Experimentalis et Applicata* **168**, 113-117 (2020).
48. H. U. Stotz.et al. Induced Plant Defense Responses against Chewing Insects. Ethylene Signaling Reduces Resistance of Arabidopsis against Egyptian Cotton Worm But Not Diamondback Moth1. *Plant Physiology* **124**, 1007-1018 (2000).
49. F. G. Horgan, A. M. Romena, C. C. Bernal, M. L. P. Almazan&A. F. Ramal. Stem borers revisited: Host resistance, tolerance, and vulnerability determine levels of field damage from a complex of Asian rice stemborers. *Crop Protection* **142**, 105513 (2021).
50. S. N. Johnson&M. Riegler. Root damage by insects reverses the effects of elevated atmospheric CO₂ on eucalypt seedlings. *Plos One* **8**, e79479 (2013).
51. E. Panel o. P. Health.et al. Pest categorisation of *Spodoptera litura*. *EFSA Journal* **17**, e05765 (2019).
52. J. W. Clohessy.et al. Development of a high-throughput plant disease symptom severity assessment tool using machine learning image analysis and integrated geolocation. *Computers and*

- Electronics in Agriculture* **184**, 106089 (2021).
53. H. Tong&Z. Nikoloski. Machine learning approaches for crop improvement: Leveraging phenotypic and genotypic big data. *Journal of plant physiology* **257**, 153354 (2021).
 54. J. Redmon, S. Divvala, R. Girshick&A. Farhadi. You Only Look Once: Unified, Real-Time Object Detection. *arXiv e-prints* arXiv:1506.02640 (2015).
 55. P. Y. Simard, D. Steinkraus&J. C. Platt. Best practices for convolutional neural networks applied to visual document analysis. *Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings.* 958-963 (2003).
 56. M. Abadi.et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. *arXiv preprint arXiv:1603.04467* (2016).
 57. A. Krizhevsky, I. Sutskever&G. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. *Advances in neural information processing systems* **25**, (2012).
 58. K. Simonyan&A. Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv e-prints* arXiv:1409.1556 (2014).
 59. K. He, X. Zhang, S. Ren&J. Sun. Deep Residual Learning for Image Recognition. *arXiv e-prints* arXiv:1512.03385 (2015).
 60. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens&Z. Wojna. Rethinking the Inception Architecture for Computer Vision. *arXiv e-prints* arXiv:1512.00567 (2015).
 61. G. Huang, Z. Liu, L. van der Maaten&K. Q. Weinberger. Densely Connected Convolutional Networks. *arXiv e-prints* arXiv:1608.06993 (2016).
 62. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever&R. Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
 63. D. P. Kingma&J. Ba. Adam: A Method for Stochastic Optimization. *arXiv e-prints* arXiv:1412.6980 (2014).
 64. S. Ruder. An overview of gradient descent optimization algorithms. *arXiv e-prints* arXiv:1609.04747 (2016).
 65. K. Weiss, T. M. Khoshgoftaar&D. Wang. A survey of transfer learning. *Journal of Big data* **3**, 1-40 (2016).
 66. N. Ketkar&N. Ketkar. Introduction to keras. *Deep learning with python: a hands-on introduction* 97-111 (2017).
 67. S. Lathuilière, P. Mesejo, X. Alameda-Pineda&R. Horaud. A Comprehensive Analysis of Deep Regression. *arXiv e-prints* arXiv:1803.08450 (2018).
 68. A. Krizhevsky, I. Sutskever&G. E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
 69. X. Liu, M. Chi, Y. Zhang&Y. Qin. Classifying High Resolution Remote Sensing Images by Fine-Tuned VGG Deep Networks. *IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium* 7137-7140 (2018).
 70. K. Pearson. Note on Regression and Inheritance in the Case of Two Parents. *Proceedings of the Royal Society of London Series I* **58**, 240-242 (1895).
 71. X. Shi.et al. The complete reference genome for grapevine (*Vitis vinifera* L.) genetics and breeding. *Horticulture Research* **10**, uhad061 (2023).
 72. X. Zhou&M. Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nature genetics* **44**, 821-824 (2012).
 73. R. Leinonen, H. Sugawara, M. Shumway&I. N. S. D. Collaboration. The sequence read archive.

- 917 *Nucleic acids research* **39**, D19-D21 (2010).
- 918 74. A. M. Bolger, M. Lohse & B. Usadel. Trimmomatic: a flexible trimmer for Illumina sequence data.
- 919 *Bioinformatics* **30**, 2114-2120 (2014).
- 920 75. M. Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads.
- 921 *EMBnet. journal* **17**, 10-12 (2011).
- 922 76. A. Dobin et al. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21 (2013).
- 923 77. M. I. Love, W. Huber & S. Anders. Moderated estimation of fold change and dispersion for RNA-
- 924 seq data with DESeq2. *Genome biology* **15**, 1-21 (2014).

ACCEPTED MANUSCRIPT