









## REVIEW ARTICLE

# Genomics-based plant disease resistance prediction using machine learning

Shriprabha R. Upadhyaya<sup>1,2</sup>  | Monica F. Danilevich<sup>1,2</sup>  | Aria Dolatabadian<sup>1</sup>  |  
Ting Xiang Neik<sup>3</sup>  | Fangning Zhang<sup>4</sup> | Hawlader A. Al-Mamun<sup>1,2</sup>  |  
Mohammed Bennamoun<sup>5</sup>  | Jacqueline Batley<sup>1</sup>  | David Edwards<sup>1,2</sup> 

<sup>1</sup>School of Biological Sciences, The University of Western Australia, Perth, Western Australia, Australia

<sup>2</sup>Centre for Applied Bioinformatics, The University of Western Australia, Perth, Western Australia, Australia

<sup>3</sup>NUS Agritech Centre, National University of Singapore, Singapore, Republic of Singapore

<sup>4</sup>College of Life Sciences, Shandong Normal University, Jinan, China

<sup>5</sup>School of Physics, Mathematics and Computing, University of Western Australia, Perth, Western Australia, Australia

## Correspondence

David Edwards, School of Biological Sciences, The University of Western Australia, Perth, WA 6009, Australia.  
Email: [dave.edwards@uwa.edu.au](mailto:dave.edwards@uwa.edu.au)

## Funding information

Australian Research Council, Grant/Award Number: DP200100762 and DP210100296

## Abstract

Plant disease outbreaks continuously challenge food security and sustainability. Traditional chemical methods used to treat diseases have environmental and health concerns, raising the need to enhance inherent plant disease resistance mechanisms. Traits, including disease resistance, can be linked to specific loci in the genome and identifying these markers facilitates targeted breeding approaches. Several methods, including genome-wide association studies and genomic selection, have been used to identify important markers and select varieties with desirable traits. However, these traditional approaches may not fully capture the non-linear characteristics of the effect of genomic variation on traits. Machine learning, known for its data-mining abilities, offers an opportunity to enhance the accuracy of the existing trait association approaches. It has found applications in predicting various agronomic traits across several species. However, its use in disease resistance prediction remains limited. This review highlights the potential of machine learning as a complementary tool for predicting the genetic loci contributing to pathogen resistance. We provide an overview of traditional trait prediction methods, summarize machine-learning applications, and address the challenges and opportunities associated with machine learning-based crop disease resistance prediction.

## KEYWORDS

agriculture, disease resistance, GWAS, machine learning, SNPs, trait prediction

## 1 | INTRODUCTION

Plant disease outbreaks have long posed a significant challenge to agriculture, threatening food security and causing an average global yield loss of 20%–30% (Ristaino et al., 2021). With the world's population projected to reach 10 billion people, a 60% increase in food production is required by the year 2050 (FAO, 2019;

Fedoroff, 2015; Ristaino et al., 2021). Reducing crop losses due to pests and pathogens is important to meet future demand. Traditional chemical approaches have been employed to protect plants from disease, but they come with environmental and health concerns (Brodie et al., 2012; Damalas & Eleftherohorinos, 2011; Davies et al., 1994; Kole et al., 2001). Moreover, the emergence of pesticide-resistant pathogens has further eroded the effectiveness

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2024 The Author(s). *Plant Pathology* published by John Wiley & Sons Ltd on behalf of British Society for Plant Pathology.

of these measures (Berger et al., 2017; Bolton et al., 2012; Cazorla et al., 2002; Lamichane et al., 2016; Shenge et al., 2014). A more sustainable approach to disease management in crops lies in understanding and exploiting the plant's innate disease resistance mechanisms (Toojinda et al., 2000).

There are two main types of disease resistance: quantitative and qualitative. Quantitative disease resistance is controlled by multiple genes interacting with each other and the environment to reduce disease severity, potentially influencing several other traits (Lynch & Walsh, 1998; St. Clair, 2010; Young, 1996). In contrast, qualitative disease resistance, also known as monogenic disease resistance, is determined by single genes (Flor, 1971; St. Clair, 2010). Understanding the mechanisms of plant disease resistance and the underlying genetic factors involved provides necessary information to breeders for developing strategies to breed resistant plants, thereby enhancing productivity and sustainable agriculture. Molecular breeding methods such as marker-assisted selection (MAS) can help identify individuals carrying favourable alleles that can be selected for breeding. For example, Warburton et al. (2023) identified markers and underlying metabolic pathways associated with resistance to fall armyworm in maize. Their study found that resistant maize plants typically have higher chlorophyll content than susceptible genotypes, which show lower chlorophyll *b*, lutein and  $\beta$ -carotene. The genes associated with the identified markers could be introduced into susceptible maize lines to develop resistant hybrids. Quantitative trait loci (QTLs) contribute to the overall resistance against pathogens, although their effects vary. For example, some QTLs may confer resistance to a wide range of diseases, such as the *Rpp* loci in soybean providing resistance against soybean rust caused by *Phakopsora pachyrhizi* (Akamatsu et al., 2013), while others may offer protection against specific pathogens, for example the rice blast resistance gene *Pi-ta* on chromosome 12 in rice that confers resistance against *Magnaporthe grisea* (Bryan et al., 2000; Chanchu et al., 2023). This diversity in functionality highlights the complex interplay between plant genetics and pathogen dynamics, emphasizing the need for comprehensive research to unravel the intricate mechanisms of plant disease resistance.

Molecular breeding approaches such as genomic selection have provided insights into the genetic basis of resistance and have facilitated targeted breeding efforts (Poland & Rutkoski, 2016). Genomic selection estimates breeding values for traits in individuals using genome-wide genetic markers. It allows breeders to estimate the plant's potential performance and provides the ability to select individuals with greater disease resistance (Heffner et al., 2009). A popular genomic selection method, genomic best linear unbiased prediction (gBLUP), constructs a genomic relationship matrix and combines it with phenotype data to estimate breeding values (Clark & van der Werf, 2013). In comparison, genome-wide association studies (GWAS) identify a set of genetic variations associated with specific phenotypes across a group of individuals. GWAS has been a valuable tool for characterizing the genomic basis for disease resistance, such as the identification of candidate genes related to *Rlm6* blackleg resistance in *Brassica juncea* (Yang et al., 2021). However,

these methods suffer from limitations. While genomic selection excels at identifying major-effect genes and markers, it struggles to identify rare genetic variants that may play a significant role in crop traits (Zhao et al., 2014). Moreover, traditional approaches may not fully capture the non-linear characteristics of the impact of genomic variation on traits. Despite the identification of thousands of trait-associated loci using GWAS, the causal genes at these loci often remain elusive. The number of variants identified through GWAS makes it impractical to investigate them functionally, complicating the identification of genes for functional studies (Nicholls et al., 2020).

Machine learning is an emerging field that can enhance the performance and interpretation of trait association (Hu et al., 2020). This has been made possible through technological advances that provide larger and more diverse datasets that can be applied to trait association. By analysing genetic data, machine learning can identify genes and alleles contributing towards traits, including disease resistance. It can predict the effectiveness of genes in plants that play an important role in defence against pathogens, uncovering the complex interactions between plants and pathogens (Sperschneider, 2020). This review provides an overview of disease resistance prediction in plants using machine learning, including examples of machine-learning applications for predicting agronomic traits. It also encompasses the latest advances and prospects for enhancing plant disease resistance through genomics and machine-learning methodologies.

## 2 | MACHINE LEARNING FOR THE PREDICTION OF AGRONOMIC TRAITS

Machine learning, a subset of artificial intelligence (AI), involves training algorithms on data to learn hidden patterns within the dataset, make predictions and decisions autonomously, and improve decisions over time when exposed to more data (Jordan & Mitchell, 2015; Zhang, 2020). While AI encompasses a broad range of techniques that mimic human intelligence, including reasoning and problem solving, machine learning specifically focuses on computers learning from and making decisions based on data. In developing a machine-learning model, the dataset is typically split into training and test sets. The model is trained on the training dataset to make predictions, and its performance is evaluated using the test dataset. Supervised learning models rely on labelled training datasets to learn the mapping from inputs to outputs, whereas unsupervised learning analyses unlabelled data to uncover hidden patterns (Hastie et al., 2001; Zhang, 2020). The performance and efficiency of the model are optimized by adjusting its parameters and structure, and the models' performance on the test set is evaluated using various metrics such as accuracy, precision, recall, F1-score, mean squared error (MSE) and Area Under Receiver Operating Characteristic Curve (AUROC) scores. Common machine-learning models include linear regression, random forest (RF), gradient boosting (GB), support vector machines (SVM) and neural networks.

Machine-learning algorithms are well-suited for analysing complex agronomic datasets, often involving numerous variables. These algorithms can extract patterns, learn from historical data and generate models to predict agronomic traits (van Klompenburg et al., 2020). Several recent studies have demonstrated the potential of machine learning in agricultural applications (Table 1). For example, Wang et al. (2023) introduced a deep learning method called deep neural network genomic prediction (DNNGP), outperforming traditional linear regression models and other machine-learning

methods for predicting agronomic traits using multi-omics data in plants. Dhillon et al. (2023) combined crop modelling and machine learning to improve yield prediction for winter wheat and oilseed rape, with a RF model incorporating normalized difference vegetation index and climate variables.

Any prediction model, including machine-learning models, relies on high-quality datasets for training. These datasets can be collected from diverse sources, such as field observations, weather data, remote sensing, soil nutritional profiles and genetic information

**TABLE 1** An overview of various machine learning and deep learning models used for predicting agronomic traits.

Plant	Trait predicted	Modelling methods used	Reference
Wheat	Yield, yield components and agronomic traits	Multilayer perceptron (MLP) and convolutional neural network (CNN)	Sandhu et al. (2021)
Maize and soybean	Grain yield	CNN-recurrent neural network (CNN-RNN), random forests (RF), deep fully connected neural networks (DFNN) and least absolute shrinkage and selection operator (LASSO)	Khaki et al. (2020)
Wheat	Grain yield	Deep kernel (arc-cosine kernel, AK), non-additive Gaussian kernel (GK), genomic best linear unbiased predictor (GBLUP/GB)	Crossa et al. (2019)
Wheat, rice	Growth. Yield (drought), yield (irrigated), thousand kernel weight (TKW) and days to heading. Culm diameter, culm length, culm number, grain length, grain width, grain weight, days to heading, ligule length, leaf length, leaf width, panicle length and seedling height	Elastic net, ridge regression, LASSO regression, RF, gradient boosting machines and support vector machines (SVM), with two state-of-the-art classical statistical genetics methods; GBLUP and a two-step sequential method based on linear regression	Grinberg et al. (2020)
Soybean	Flower colour, seed coat colour, pod colour, pubescence density, seed oil content, seed protein content and seed weight	XGBoost or RF	Gill et al. (2022)
Alfalfa	Fall dormancy	SVM regression, and regularization-related methods, such as LASSO and ridge regression	Zhang et al. (2023)
Wheat	Grain yield	Neural network genomic prediction (DNNGP), GBLUP, light gradient boosting machine (LightGBM), support vector regression (SVR), deep learning genomic selection (DeepGS), deep learning genome-wide association study (DLGWAS)	Wang et al. (2023)
Sunflower, wheat	Grain yield	RF and artificial neural network (ANN)	Morales and Villalobos (2023)
Common bean	Flowering time	ANN, ridge regression best linear unbiased predictor (RR-BLUP)	Rosado et al. (2020)
Chinese thuja	Flowering period	RNN, long short-term memory and gated recurrent unit	Jiao et al. (2022)
Rice	Drought resistance	Gene regulation and association network (GRAIN)	Gupta et al. (2021)
Maize	Grain yield	Bagging, decision tree, RF and ANN-MLP	Harsányi et al. (2023)
Wheat, oilseed rape	Grain yield	Light use efficiency (LUE)+RF, RF (RF1 [input: normalized difference vegetation index {NDVI}], RF2 [input: climate variables], RF3 [input: NDVI + climate variables], RF4 [input: LUE generated biomass + climate variables]), and one semi-empiric LUE mode	Dhillon et al. (2023)
Maize	Yield	Multimodal prediction using RF, XGBoost, tabular deep neural network (tab-DNN) and spectral deep neural network (sp-DNN)	Danilevicz et al. (2021)

(Bochenek & Ustrnul, 2022; Bolten et al., 2009). Pre-processing involves data cleaning, handling missing values and feature selection, which is important in optimizing model performance and minimizing bias (Maharana et al., 2022). Furthermore, dividing data into training and validation sets, coupled with techniques such as cross-validation and hold-out validation, helps validate model accuracy and addresses potential overfitting or underfitting of the model.

### 3 | MACHINE LEARNING AS A COMPLIMENTARY TOOL FOR DISEASE RESISTANCE PREDICTION

Machine learning has been applied to predict disease resistance in several plant species, as summarized in Table 2. It was proposed as a potential approach for genomic selection in breeding wheat for rust resistance (González-Camacho et al., 2018). A machine-learning approach was applied to identify genomic regions associated with brown rust resistance in sugarcane (Aono et al., 2020). Brown rust, caused by the fungal pathogen *Puccinia melanocephala*, decreases sugarcane yield, leading to substantial economic losses. A solution for eliminating the disease involves using cultivars with inherent resistance (Wang et al., 2019). Sugarcane cultivars exhibit complex genomic organization with a ploidy between 6 and 14, aneuploidy, a large genome size of 10 Gb and a high proportion of repetitive regions (Garsmeur et al., 2018). This makes it difficult for simple statistical models to capture the non-linear attributes of the dataset, so machine-learning methods were used to identify resistance genes. In an analysis of parents and 180 progeny, the study evaluated eight machine-learning algorithms, including K-Nearest Neighbour (KNN; Cover & Hart, 1967), SVM (Cortes & Vapnik, 1995), Gaussian Process (GP; Rasmussen & Williams, 2006), Decision Tree (DT; Quinlan, 1986), RF (Breiman, 2001), Multilayer Perceptron (MLP) neural network (Murtagh, 1991), Adaptive Boosting (AB; Freund & Schapire, 1997) and Gaussian Naive Bayes (GNB; Friedman et al., 1997). Five different Feature Selection (FS) methods were used to reduce the marker dataset by obtaining feature importance. The resulting dataset comprised 131 single-nucleotide polymorphisms (SNPs) and an accuracy of 95%. Many identified regions corresponded to previously identified QTLs (Aono et al., 2020).

Similar studies were conducted on disease resistance prediction in sugarcane by Pimenta et al. (2021). Sugarcane yellow leaf (SCYL) caused by sugarcane yellow leaf virus (SCYLV) is a major disease that impairs plant development, and breeding for plant cultivars that are resistant to SCYL is a promising approach to mitigate the loss and damage caused by the virus. Several genome-wide studies were assessed for efficacy in identifying the markers and genes associated with SCYLV resistance. GWAS methods such as FarmCPU, mixed linear modelling and machine-learning algorithms coupled with feature selection were used to predict genotypes associated with resistance. The genotype dataset consisted of amplified fragment length polymorphisms (AFLPs), simple-sequence repeats (SSRs), SNPs and insertions and deletions (indels). Eight different machine-learning

algorithms were tested using the full marker set, resulting in a low accuracy; however, a combination of feature selection with MLP improved accuracy to 95%. The study was consistent with Aono et al. (2020), where a reduced marker set of 120–190 SNPs yielded higher accuracy than a larger dataset. Several markers associated with disease resistance matched those identified using FarmCPU and mixed modelling analyses (Pimenta et al., 2021). In another study, a QTL controlling resistance to sugarcane mosaic virus (SCMV), largely unexplored in sugarcane, was predicted using GWAS and machine learning coupled with feature selection. Using a diverse panel of 97 genotypes, GWAS was conducted using a mixed linear modelling method and eight machine-learning models, combined with three feature selection methods to predict susceptible and resistant genotypes. Consistent with previous studies, a reduced dataset of 73 SNPs resulted in a mean accuracy of 90.2% for different machine-learning algorithms, with MLP having the highest accuracy of 99.7% (Pimenta et al., 2023). They extended this study by validating the markers obtained from machine learning and annotating them using RNA-Seq data to identify the biological processes involved in resistance (Pimenta et al., 2023).

*Hemileia vastatrix* causing coffee leaf rust is the main fungal disease that impacts the worldwide cultivation of coffee. A total of 245 Arabica coffee plants were genotyped with 137 markers. The study used decision tree models with refinements such as bagging, RF and boosting to predict the resistance to coffee leaf rust. They compared the method with an artificial neural network (ANN) and a genomic selection method known as Bayesian Generalized Linear Regression (BGLR; Pérez & de los Campos, 2014). An average accuracy of 30.6%–60.7% was obtained, outperforming the BGLR method. The top 10% of the most important markers identified by each method were chosen and compared with results in the literature. A total of 11 markers were presented in common between decision tree with boosting and statistical Generalized Bayesian Lasso (GBLASSO) methods. A comparison with the literature indicated an average of 9.3 predicted markers were present in regions of QTLs associated with coffee leaf rust resistance. The study concluded that the machine-learning models achieved higher accuracy overall, and the authors identified more markers associated with the trait compared to the GBLASSO method (Sousa et al., 2021).

Two machine-learning models, MLP and probabilistic neural network (PNN), were compared using maize and wheat high-throughput molecular markers. Trait combinations, including maize grey leaf spot (GLS) resistance, were evaluated. (González-Camacho et al., 2016). Sixteen genomic datasets were evaluated, of which six contained information on GLS resistance. AUROC score, an evaluation metric used to identify how well a model can distinguish between classes, revealed that PNN had consistently better scores of 0.05 higher than MLP, demonstrating better accuracy in classifying the GLS resistance phenotype (González-Camacho et al., 2016).

Several studies use a combination of statistical and machine-learning models for genomic prediction of disease resistance. For example, González-Camacho et al. (2012) employed linear Bayesian least absolute shrinkage and selection operator (LASSO)

TABLE 2 Summary of machine-learning models used for disease resistance prediction in plants.

Disease	Plant	Type of model	Dataset	Evaluation metrics	Reference
Brown rust	Sugarcane	KNN, SVM, GP, DT, RF, MLP, AB and GNB	SNP data	Accuracy 95%	Aono et al. (2020)
Sugarcane yellow leaf	Sugarcane	KNN, SVM, GP, DT, RF, MLP, AB and GNB	A panel of markers	Accuracy 95%	Pimenta et al. (2021)
Orange rust	Coffee	ANN, DT with refinements	Marker set	Accuracy 30.6%–60.7%	Sousa et al. (2021)
Grey leaf spot	Maize	MLP and PNN	SNP chips	Area under curve	González-Camacho et al. (2016)
Grey leaf spot	Maize	RKHS and RBFNN	Marker set	Correlation –0.25 to 0.59	González-Camacho et al. (2012)
Northern corn leaf blight	Maize	RKHS and RBFNN	Marker set	Correlation –0.4 to 0.7	González-Camacho et al. (2012)
Sugarcane mosaic virus	Sugarcane	KNN, SVM, GP, DT, RF, MLP, AB and GNB	SNP set	Accuracy 92%–97%	Pimenta et al. (2023)
Bacterial blight	Rice	ANN	Protein sequence	Matthews correlation coefficient –0.44	Xia et al. (2009)

Abbreviations: AB, adaptive boosting; ANN, artificial neural network; DT, decision tree; GNB, Gaussian naive Bayes; GP, Gaussian process; KNN, K-nearest neighbour; MLP, multilayer perceptron neural network; PNN, probabilistic neural network; RBFNN, radial basis function neural networks; RF, random forest; RKHS, reproducing kernel Hilbert spaces; SNP, single-nucleotide polymorphism; SVM, support vector machine.

regression and two non-linear machine-learning models, reproducing kernel Hilbert spaces (RKHS) regression and radial basis function neural networks (RBFNN), to study simulated and real maize genotyping data representing 55,000 markers. They assessed 300 maize lines for resistance to diseases such as GLS caused by *Cercospora zea-maydis* and northern corn leaf blight caused by *Exserohilum turcicum*, alongside other agronomic traits. The study revealed that machine-learning models exhibited a higher correlation for resistance prediction, indicating their ability to capture epistatic patterns within the marker set (González-Camacho et al., 2012). Another study favoured Support Vector Classification with linear kernel (SVC-lin) among three classifiers and six other regressor models that were evaluated for predicting disease resistance for GLS and stem rust across 14 maize and 16 wheat datasets (Ornella et al., 2014). A study conducted by Ornella et al. (2012) showed statistical methods to have a higher prediction accuracy compared to machine-learning models for genomic prediction of resistance to stem rust (*Puccinia graminis*) and yellow rust (*Puccinia striiformis*). Bayesian LASSO (BL), Ridge Regression (RR) and Support Vector Regression (SVR) with linear or radial basis function (RBF) kernel models were evaluated using five CIMMYT wheat populations, with linear statistical models BL and RR demonstrating better prediction accuracy than SVR due to the additive effects of the two traits (Ornella et al., 2012).

An ANN has been applied to identify disease resistance genes in rice, which provides resistance against bacterial blight caused by *Xanthomonas oryzae* pv. *oryzae* (Xoo). A protein sequence dataset was used to train a back propagation neural network model to identify candidate Xoo-resistant genes (Xia et al., 2009). These techniques extend to predicting disease resistance (R) proteins based on amino acid sequences in various plants using methods such as RF and SVM (Kushwaha et al., 2015; Pal et al., 2016; Simón et al., 2022).

In another study, a plant R protein predictor called prPred was developed based on a SVM that can distinguish plant R proteins from other proteins (Wang, Wang, et al., 2021).

#### 4 | A COMPARISON OF MACHINE-LEARNING METHODS FOR TRAIT ASSOCIATION

Machine-learning algorithms have successfully identified genes associated with disease resistance and susceptibility across several crops. Most studies based on genomic selection use supervised machine-learning algorithms as they have the advantage of using labelled datasets to draw the relationship between the observations, which provides the model with examples of desirable phenotypes. The supervised methods use these examples to evaluate whether its prediction was correct, subsequently adjusting its internal weights or tree splitting to improve prediction accuracy. For example, Aono et al. (2020) employed a SNP dataset from 219 sugarcane individuals and their observed disease resistance rankings based on percentage of leaf area infected to teach multiple machine-learning models to identify genomic regions associated with sugarcane brown rust resistance. Knowing the phenotypic variation between individuals in the dataset enables the machine-learning model to autonomously rank discernible features associated with the desired phenotype. Nonetheless, each machine-learning model operates differently, requiring some consideration when choosing the algorithm, depending on the data availability, analysis objective and explainability required. Traditional machine-learning algorithms have diverse assumptions about the dataset; for example, SVM assumes that data is linearly separable. SVM and Logistic Regression use hyperplanes to find the maximum difference between the classes within a dataset (Cortes &



Vapnik, 1995; Hosmer Jr et al., 2013). KNNs are instance-based algorithms that use proximity, where a new data point is predicted based on the average value of its 'K' nearest neighbour or similar cases that are available (Cover & Hart, 1967). GPs are probabilistic algorithms that model the underlying function of the data and predict outcomes by assuming that the data points in input follow a joint Gaussian distribution (Rasmussen & Williams, 2006). While KNNs are flexible and require no training phase, GPs require a computationally intensive training phase. However, GPs offer the advantage of computing an associated uncertainty value along with predictions, making it helpful in understanding confidence levels. In contrast, RF is an ensemble algorithm powered by aggregating the predictions of multiple individual decision trees trained on a random subset of the data. Each decision tree prediction is combined to provide a result based on the majority agreement, which reduces the risk of overfitting and bias due to variability within the decision tree cohort (Breiman, 2001). Similarly, XGBoost is a scalable decision tree system, optimized for handling sparse data and with the option to be parallelized across multiple machines to accelerate training (Tianqi & Carlos, 2016). RF and XGBoost are increasingly popular methods for approaching regression and classification tasks in tabular datasets, outperforming other algorithms across multiple studies (Gill et al., 2022; Xu et al., 2021).

No single algorithm will consistently excel in all kinds of problems. This idea was postulated in the No-Free-Lunch theorem, which states that all optimization strategies perform equally well when averaged across all possible problems, meaning researchers should compare multiple algorithms to find the best fit for the phenomenon they want the model to predict (Goldblum et al., 2023). This concept is demonstrated in multiple studies using machine-learning algorithms to identify disease-related markers. For example, González-Camacho et al. (2018) compared reproducing kernel Hilbert space, Bayesian Lasso, Ridge Regression and SVM, finding that the latter provided the best classification for wheat rust resistance based on 16 SNP datasets. Another study focussing on the prediction of genetic values for wheat rust resistance compared eight machine-learning models, finding that Bayesian Lasso had superior predictive power (Ornella et al., 2012). Aono et al. (2020) observed that Gaussian Process, Multilayer Perceptron and Gaussian Naive Bayes alternated in being the most accurate model depending on the SNP dataset employed for predicting brown rust phenotypic groups. Pimenta et al. (2023) compared eight machine-learning algorithms for a multi-omics investigation of sugarcane mosaic virus resistance, finding that RF had a higher performance (66.9% accuracy) when using the full dataset, but was outperformed by the Multilayer Perceptron model after feature selection was used to reduce the SNP dataset size. In another study, Multilayer Perceptron models were outperformed by PNNs when classifying GLS resistance in maize based on area under curve (AUC) criteria (González-Camacho et al., 2016).

Neural networks are advantageous when dealing with complex traits due to their capacity to successfully model complex dependencies between features and integrate multidimensional datasets into a single prediction (Danilevicz et al., 2021; Montesinos-López et al., 2023). Multilayer Perceptron networks are a popular modality of

neural networks, and are classified as feedforward architectures with multiple layers of interconnected neurons that forward the information from the input layer to the output layer (Murtagh, 1991). MLP is a shallow neural network architecture, meaning it uses fewer layers than their deep neural network counterparts and is generally faster to train. Popular multilayered feedforward neural networks are PNNs widely used for image recognition and classification. They work by storing representations of patterns found in the input layer during training into a pattern layer. During testing, the pattern in input is compared to the stored patterns and their similarity is calculated, with the output layer computing the probability of each class. They can work with smaller training datasets; however, they may require more memory if the dataset is bigger (Mohebbi et al., 2020). More recently, deep neural networks have occupied the forefront of machine-learning development as novel layer functions enabled substantially increased neural network depth (He et al., 2016). The most popular deep neural networks are convolutional neural networks that effectively capture spatial relationships and have been widely applied to extract features from image datasets (Barbosa et al., 2020; Kattenborn et al., 2021; Wang, Zhang, et al., 2021). Long Short-Term Memory networks are well-suited to recognize temporal dependencies in time-series and sequential datasets (Graves, 2012; Kim et al., 2017; Maldonado et al., 2020; Xiao et al., 2018), and more recently, Transformer architectures have been applied that can model long-range dependencies across features and are less prone to overfitting when dealing with small datasets. Neural networks can be easily modified to build bespoke models responsive to the specific requirements of the dataset or task at hand. For example, multiple studies leveraged the flexibility of deep neural network architectures to include statistical principles to represent different aspects of genomic interaction. A Poisson deep neural network was proposed in which the Poisson distribution was used as the loss function to more accurately model 'count' data for the multivariate and univariate trait prediction of crop trait phenotypes (Montesinos-López et al., 2021, 2020). A similar deep-learning architecture using Bayesian modelling of Poisson distribution was also proposed by Rodrigo and Tsokos (2020), while another study proposed a Bayesian regularized neural network for predicting traits in maize and eucalyptus crops (Maldonado et al., 2020). Using Bayesian regularization adds a prior distribution over the model parameters, which encourages a reduction in model complexity and overfitting; it is particularly useful for understanding the uncertainty associated with the model prediction (Pearce et al., 2020). As shown here, the wide range of neural network architectures, coupled with their adaptability, holds great potential for addressing the current challenges in modelling crop disease resistance and identifying genes associated with plant susceptibility.

## 5 | CHALLENGES AND OPPORTUNITIES FOR MACHINE LEARNING-BASED CROP DISEASE RESISTANCE PREDICTION

AI, specifically machine learning, has much potential for disease resistance prediction. However, several challenges must be addressed

to fully exploit the effectiveness of machine learning-based crop disease resistance prediction, as summarized in Figure 1. These challenges include data missingness, imbalanced datasets, model interpretability and the extraction of representative datasets from large datasets.

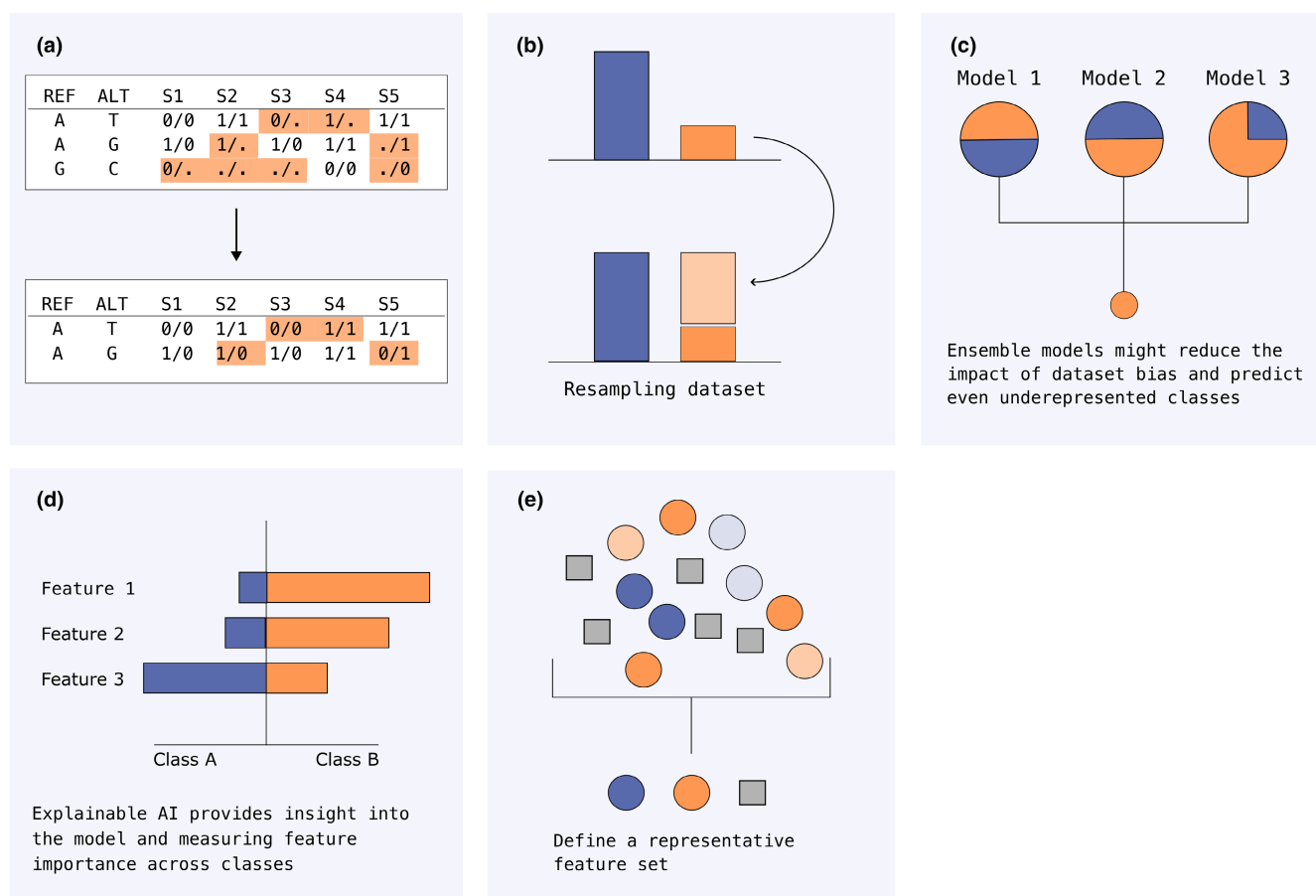
## 5.1 | Data quality

One of the primary challenges of machine learning is data quality, particularly missingness in datasets that can affect the accuracy and reliability of the prediction models (Emmanuel et al., 2021; Madhu & Rajinikanth, 2012). Several approaches, such as imputation and deletion of data, can help mitigate missingness in data. Imputation techniques work by filling in the missing values based on patterns observed in the datasets. Several imputation techniques, such as simple imputation, advanced imputation and multiple imputation, are available and can be used depending on the nature of missingness. These techniques use different statistical and machine-learning approaches such as mean, median, k-nearest neighbours

or matrix completion methods to handle missing data (Baraldi & Enders, 2010). Deletion works by removing the variables with missing values when the missingness is minimal and does not impact the overall dataset (Emmanuel et al., 2021). Several software packages and libraries are available for data imputation in machine learning (Platias & Petasis, 2020), including Python libraries such as scikit-learn (Pedregosa et al., 2018) and fancyImpute, which offers SimpleImputer, SoftImpute and BiScalar algorithms to impute and complete missing data (Rubinstein & Feldman, 2016). Missingness is often observed in SNP datasets. Packages such as BEAGLE (Browning et al., 2018) and EAGLE (Loh et al., 2016) can phase and impute SNPs to handle missingness in these datasets.

## 5.2 | Data imbalance

Imbalanced datasets are another common issue in machine learning, where the classes are unequal. The number of variables can sometimes be significantly larger than the number of samples, as explained by Thabtah et al. (2020) and Buda et al. (2018). In



**FIGURE 1** An illustration of the challenges and solutions for machine-learning methods. (a) Representation of a genomic variation file with missing data points imputed and deletion of lines with over 50% data missing. (b) Demonstration of the effect of using data resampling to reduce class representation imbalance in the dataset. (c) Shows how a combination of diverse models can reduce dataset biases, enabling the prediction of all classes. (d) Visualization of feature importance analysis to show top features influencing the model prediction. (e) Feature selection might reduce noise and redundancy in the dataset, allowing the model to focus on meaningful features.

genomic datasets such as SNP data, the number of variables (SNPs) often exceeds 1,000,000, largely outnumbering the number of samples, which may be around 200–1000. This imbalance can introduce classification bias favouring the majority class and adversely affect the performance of deep-learning models, including neural networks (Buda et al., 2018) and machine-learning classifiers (Blagus & Lusa, 2010). Several methods have been proposed to solve imbalanced datasets, including resampling techniques such as over- and under-sampling, and ensemble-based methods (Abd Elrahman & Abraham, 2013; Thabtah et al., 2020). Under-sampling is a technique to randomly eliminate instances from the majority class to balance the dataset. However, it should be used cautiously as under-sampling can lead to the loss of valuable information. Over-sampling is where the minority class is randomly replicated until it matches the majority class (Fernández et al., 2018). Techniques such as Synthetic Minority Over-sampling Technique (SMOTE) and Adaptive Synthetic (ADASYN) can generate synthetic samples to balance datasets rather than replicating the dataset (Blagus & Lusa, 2013; Haibo et al., 2008). In ensemble-based methods, several classifiers are trained and their decisions are combined into a single class label to improve accuracy (Rokach, 2010). In the study conducted by Sousa et al. (2021) to predict disease resistance in coffee, ensemble methods such as bagging and boosting were used to overcome the non-normality of phenotype values.

### 5.3 | Model interpretability

Machine learning and deep-learning models are often perceived as black boxes due to their complex architectures and numerous parameters. A common issue with deep-learning models is the lack of understanding of how and why the model makes decisions. Explainable AI (XAI) emerges as a solution by providing insights into the decision-making process of the models, overcoming the challenge of interpretability (Gunning et al., 2019). XAI enhances the model's interpretability by helping the users understand the reasoning behind the model's predictions by highlighting the features or attributes that influence these predictions (Barredo Arrieta et al., 2020; Gunning et al., 2019). Such an approach aids in identifying potential bias in the model and translates the model's predictions into biological insights. XAI methods such as SHapley Additive exPlanations (SHAP; Lundberg & Lee, 2017; Ribeiro et al., 2016) and feature permutation methods have been used to identify and extract feature importance and their contributions towards model predictions in several genomic and phenotypic studies (Bayer, Scheben, et al., 2021; Danilevicz et al., 2021; Upadhyaya et al., 2022).

### 5.4 | Large datasets

Handling vast amounts of data is a recurrent challenge in machine learning as large datasets can contain irrelevant information

leading to increased computational time and poor performance (Pudjihartono et al., 2022). To overcome this challenge, feature selection and reduction techniques have been applied. Feature selection methods identify the most informative and relevant SNPs associated with trait prediction, particularly with SNP datasets. Several studies used machine-learning techniques coupled with feature selection to predict disease resistance traits, achieving a prediction accuracy of up to 95% and reducing the SNP marker data size to 200 SNPs (Aono et al., 2020; Pimenta et al., 2021). The reduction in dataset size demonstrated consistently higher accuracy levels of over 90% across multiple studies (Aono et al., 2020; Pimenta et al., 2021, 2023). Methods such as GWAS and LD pruning can also help filter the datasets to improve the efficiency and accuracy of the models.

One advantage of machine learning is the potential to reuse previously under-utilized datasets. This enables researchers to reanalyse and enhance the findings obtained from prior studies or discover hidden patterns not initially explored. For example, soybean accessions from the United States Department of Agriculture (USDA) soybean germplasm collection, containing accessions that have previously been evaluated in various studies, were reused for genomic-based prediction of agronomic traits using several machine-learning and deep-learning models (Gill et al., 2022). Other machine-learning studies, such as the identification of long non-coding RNAs in plant genomes (Danilevicz et al., 2023), assessing gene models to improve gene annotation (Upadhyaya et al., 2022) and predicting phenotypes in wheat and rice (Grinberg et al., 2020), have adopted this approach of repurposing previously examined datasets. This approach promotes efficient data use and maximizes the knowledge gained from existing resources.

While most recent studies have relied on SNPs and single reference genomes for the genetic mapping of disease resistance genes, a better approach is to use pangenomes that include structural variation (SV) such as copy number variation (CNV) and presence/absence variation (PAV) for disease resistance identification (Bayer, Petereit, et al., 2021; Danilevicz et al., 2020; Golicz et al., 2020). Pangenome encompasses a whole set of genes present in the species, including the core and variable genes, and is a powerful approach to understanding SVs within species (Hurgobin & Edwards, 2017). Pangenomes are particularly relevant as disease resistance genes frequently show presence/absence variation in crop species (Golicz, Batley, & Edwards, 2016; Golicz, Bayer, et al., 2016). SV studies have now been linked with plant defence-related mechanisms (Dolatabadian et al., 2017). An approach using machine learning with pangenome datasets may provide more accurate predictions than using single reference genomes (Edwards & Batley, 2022).

## 6 | CONCLUSION

Disease resistance plays an important role in crop production, and understanding the genetic factors involved in plant disease resistance is important for developing strategies to combat pathogens and



breed resilient cultivars. While genomic selection and GWAS have been used extensively to predict disease resistance loci in plants, the causal genes at these loci often remain elusive. Machine learning offers a promising approach to uncovering hidden patterns underlying genetic traits and predicting plant disease resistance. While its applications have extended to predicting various agronomic traits such as yield, growth and flowering time, the prediction of disease resistance remains mostly unexplored. Some examples demonstrate the potential of machine-learning models for predicting disease resistance genes in crops such as rice, wheat, maize and sugarcane, but it is worth noting that a single standalone method cannot be universally recommended as the best, highlighting the need for a diversified approach based on specific contexts. Although machine learning comes with various data challenges, it also offers the potential to reuse existing datasets and expand to more complete genome references such as pangenomes. Such approaches can promote efficient data use and maximize the knowledge gained from existing resources.

## ACKNOWLEDGEMENTS

This work was supported by funding from the Australian Research Council projects DP210100296 and DP200100762. SU was supported by a partial postgraduate scholarship and a scholarship for International Research Fees from the University of Western Australia.

## DATA AVAILABILITY STATEMENT

Data sharing is not applicable—no new data is generated, or the article describes entirely theoretical research.

## ORCID

Shriprabha R. Upadhyaya  <https://orcid.org/0000-0001-9511-9562>

Monica F. Danilevicz  <https://orcid.org/0000-0001-7599-8184>

Aria Dolatabadian  <https://orcid.org/0000-0002-2158-4485>

Ting Xiang Neik  <https://orcid.org/0000-0002-2816-0458>

Hawlder A. Al-Mamun  <https://orcid.org/0000-0003-2453-0914>

Mohammed Bennamoun  <https://orcid.org/0000-0002-6603-3257>

Jacqueline Batley  <https://orcid.org/0000-0002-5391-5824>

David Edwards  <https://orcid.org/0000-0001-7599-6760>

## REFERENCES

Abd Elrahman, S.M. & Abraham, A. (2013) A review of class imbalance problem. *Journal of Network and Innovative Computing*, 1, 332–340.

Akamatsu, H., Yamanaka, N., Yamaoka, Y., Soares, R.M., Morel, W., Ivancovich, A.J.G. et al. (2013) Pathogenic diversity of soybean rust in Argentina, Brazil, and Paraguay. *Journal of General Plant Pathology*, 79, 28–40.

Aono, A.H., Costa, E.A., Rody, H.V.S., Nagai, J.S., Pimenta, R.J.G., Mancini, M.C. et al. (2020) Machine learning approaches reveal genomic regions associated with sugarcane brown rust resistance. *Scientific Reports*, 10, 20057.

Baraldi, A.N. & Enders, C.K. (2010) An introduction to modern missing data analyses. *Journal of School Psychology*, 48, 5–37.

Barbosa, A., Trevisan, R., Hovakimyan, N. & Martin, N.F. (2020) Modeling yield response to crop management using convolutional neural networks. *Computers and Electronics in Agriculture*, 170, 105197.

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bénéttot, A., Tabik, S., Barbado, A. et al. (2020) Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115.

Bayer, P.E., Petereit, J., Danilevicz, M.F., Anderson, R., Batley, J. & Edwards, D. (2021) The application of pangenomics and machine learning in genomic selection in plants. *Plant Genome*, 14, e20112.

Bayer, P.E., Scheben, A., Golicz, A.A., Yuan, Y., Faure, S., Lee, H. et al. (2021) Modelling of gene loss propensity in the pangenomes of three *Brassica* species suggests different mechanisms between polyploids and diploids. *Plant Biotechnology Journal*, 19, 2488–2500.

Berger, S., El Chazli, Y., Babu, A.F. & Coste, A.T. (2017) Azole resistance in *Aspergillus fumigatus*: a consequence of antifungal use in agriculture? *Frontiers in Microbiology*, 8, 1024.

Blagus, R. & Lusa, L. (2010) Class prediction for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 11, 523.

Blagus, R. & Lusa, L. (2013) SMOTE for high-dimensional class-imbalanced data. *BMC Bioinformatics*, 14, 106.

Bochenek, B. & Ustrnul, Z. (2022) Machine learning in weather prediction and climate analyses—applications and perspectives. *Atmosphere*, 13, 180.

Bolten, J.D., Crow, W.T., Zhan, X., Jackson, T.J. & Reynolds, C.A. (2009) Evaluating the utility of remotely sensed soil moisture retrievals for operational agricultural drought monitoring. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 3, 57–66.

Bolton, M.D., Rivera-Varas, V., del Río Mendoza, L.E., Khan, M.F.R. & Secor, G.A. (2012) Efficacy of variable tetraconazole rates against *Cercospora beticola* isolates with differing in vitro sensitivities to DMI fungicides. *Plant Disease*, 96, 1749–1756.

Breiman, L. (2001) Random forests. *Machine Learning*, 45, 5–32.

Brodie, J.E., Kroon, F.J., Schaffelke, B., Wolanski, E.C., Lewis, S.E., Devlin, M.J. et al. (2012) Terrestrial pollutant runoff to the great barrier reef: an update of issues, priorities and management responses. *Marine Pollution Bulletin*, 65, 81–100.

Browning, B.L., Zhou, Y. & Browning, S.R. (2018) A one-penny imputed genome from next-generation reference panels. *The American Journal of Human Genetics*, 103, 338–348.

Bryan, G.T., Wu, K.S., Farrall, L., Jia, Y., Hershey, H.P., McAdams, S.A. et al. (2000) A single amino acid difference distinguishes resistant and susceptible alleles of the rice blast resistance gene *Pi-ta*. *The Plant Cell*, 12, 2033–2046.

Buda, M., Maki, A. & Mazurkowski, M.A. (2018) A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, 106, 249–259.

Cazorla, F.M., Arrebola, E., Sesma, A., Pérez-García, A., Codina, J.C., Murillo, J. et al. (2002) Copper resistance in *Pseudomonas syringae* strains isolated from mango is encoded mainly by plasmids. *Phytopathology*, 92, 909–916.

Chanchu, T., Yimram, T., Chankaew, S., Kaga, A. & Somta, P. (2023) Mapping QTLs controlling soybean rust disease resistance in Chiang Mai 5, an induced mutant cultivar. *Genes*, 14, 19.

Clark, S.A. & van der Werf, J. (2013) Genomic best linear unbiased prediction (gBLUP) for the estimation of genomic breeding values. *Methods in Molecular Biology*, 1019, 321–330.

Cortes, C. & Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273–297.

Cover, T. & Hart, P. (1967) Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13, 21–27.

Crossa, J., Martini, J.W.R., Gianola, D., Pérez-Rodríguez, P., Jarquin, D., Juliana, P. et al. (2019) Deep kernel and deep learning for genome-based prediction of single traits in multi-environment breeding trials. *Frontiers in Genetics*, 10, 1168.

- Damalas, C.A. & Eleftherohorinos, I.G. (2011) Pesticide exposure, safety issues, and risk assessment indicators. *International Journal of Environmental Research and Public Health*, 8, 1402–1419.
- Danilevicz, M.F., Bayer, P.E., Boussaid, F., Bennamoun, M. & Edwards, D. (2021) Maize yield prediction at an early developmental stage using multispectral images and genotype data for preliminary hybrid selection. *Remote Sensing*, 13, 3976.
- Danilevicz, M.F., Gill, M., Fernandez, C.G.T., Petereit, J., Upadhyaya, S.R., Batley, J. et al. (2023) DNABERT-based explainable lncRNA identification in plant genome assemblies. *Computational and Structural Biotechnology Journal*, 21, 5676–5685.
- Danilevicz, M.F., Tay Fernandez, C.G., Marsh, J.I., Bayer, P.E. & Edwards, D. (2020) Plant pangenomics: approaches, applications and advancements. *Current Opinion in Plant Biology*, 54, 18–25.
- Davies, P., Cook, L. & Barton, J. (1994) Triazine herbicide contamination of Tasmanian streams: sources, concentrations and effects on biota. *Marine and Freshwater Research*, 45, 209.
- Dhillon, M.S., Dahms, T., Kuebert-Flock, C., Rummler, T., Arnault, J., Steffan-Dewenter, I. et al. (2023) Integrating random forest and crop modeling improves the crop yield prediction of winter wheat and oil seed rape. *Frontiers in Remote Sensing*, 3, 1010978.
- Dolatabadian, A., Patel, D.A., Edwards, D. & Batley, J. (2017) Copy number variation and disease resistance in plants. *Theoretical and Applied Genetics*, 130, 2479–2490.
- Edwards, D. & Batley, J. (2022) Graph pangenomes find missing heritability. *Nature Genetics*, 54, 919–920.
- Emmanuel, T., Maupong, T., Mpoeleng, D., Semong, T., Mphago, B. & Tabona, O. (2021) A survey on missing data in machine learning. *Journal of Big Data*, 8, 140.
- FAO. (2019) *The state of food and agriculture 2019. Moving forward on food loss and waste reduction*. Rome: FAO.
- Fedoroff, N.V. (2015) Food in a future of 10 billion. *Agriculture & Food Security*, 4, 11.
- Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B. & Herrera, F. (2018) *Learning from imbalanced data sets*. Cham: Springer International Publishing.
- Flor, H.H. (1971) Current status of the gene-for-gene concept. *Annual Review of Phytopathology*, 9, 275–296.
- Freund, Y. & Schapire, R.E. (1997) A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.
- Friedman, N., Geiger, D. & Goldszmidt, M. (1997) Bayesian network classifiers. *Machine Learning*, 29, 131–163.
- Garsmeur, O., Droc, G., Antonise, R., Grimwood, J., Potier, B., Aitken, K. et al. (2018) A mosaic monoploid reference sequence for the highly complex genome of sugarcane. *Nature Communications*, 9, 2638.
- Gill, M., Anderson, R., Hu, H., Bennamoun, M., Petereit, J., Valliyodan, B. et al. (2022) Machine learning models outperform deep learning models, provide interpretation and facilitate feature selection for soybean trait prediction. *BMC Plant Biology*, 22, 180.
- Goldblum, M., Finzi, M., Rowan, K. & Wilson, A.G. (2023) The No free lunch theorem, Kolmogorov complexity, and the role of inductive biases in machine learning. *arXiv*, 2304.05366 [preprint].
- Golicz, A.A., Batley, J. & Edwards, D. (2016) Towards plant pangenomics. *Plant Biotechnology Journal*, 14, 1099–1105.
- Golicz, A.A., Bayer, P.E., Barker, G.C., Edger, P.P., Kim, H., Martinez, P.A. et al. (2016) The pangenome of an agronomically important crop plant *Brassica oleracea*. *Nature Communications*, 7, 13390.
- Golicz, A.A., Bayer, P.E., Bhalla, P.L., Batley, J. & Edwards, D. (2020) Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics*, 36, 132–145.
- González-Camacho, J.M., Crossa, J., Pérez-Rodríguez, P., Ornella, L. & Gianola, D. (2016) Genome-enabled prediction using probabilistic neural network classifiers. *BMC Genomics*, 17, 208.
- González-Camacho, J.M., de Los Campos, G., Pérez, P., Gianola, D., Cairns, J.E., Mahuku, G. et al. (2012) Genome-enabled prediction of genetic values using radial basis function neural networks. *Theoretical and Applied Genetics*, 125, 759–771.
- González-Camacho, J.M., Ornella, L., Pérez-Rodríguez, P., Gianola, D., Dreisigacker, S. & Crossa, J. (2018) Applications of machine learning methods to genomic selection in breeding wheat for rust resistance. *The Plant Genome*, 11, 170104.
- Graves, A. (2012) Long short-term memory. In: Graves, A. (Ed.) *Supervised sequence labelling with recurrent neural networks*. Berlin, Heidelberg: Springer, pp. 37–45.
- Grinberg, N.F., Orhobor, O.I. & King, R.D. (2020) An evaluation of machine-learning for predicting phenotype: studies in yeast, rice, and wheat. *Machine Learning*, 109, 251–277.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S. & Yang, G.-Z. (2019) XAI—Explainable artificial intelligence. *Science robotics*, 4, eaay7120.
- Gupta, C., Ramegowda, V., Basu, S. & Pereira, A. (2021) Using network-based machine learning to predict transcription factors involved in drought resistance. *Frontiers in Genetics*, 12, 652189.
- Haibo, H., Yang, B., Garcia, E.A. & Shutao, L. (2008) ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*. Hong Kong: IEEE (Institute of Electrical and Electronics Engineers), pp. 1322–1328.
- Harsányi, E., Bashir, B., Arshad, S., Ocwa, A., Vad, A., Alsalmán, A. et al. (2023) Data mining and machine learning algorithms for optimizing maize yield forecasting in central Europe. *Agronomy*, 13, 1297.
- Hastie, T., Friedman, J. & Tibshirani, R. (2001) Overview of supervised learning. In: Hastie, T., Tibshirani, R. & Friedman, J. (Eds.) *The elements of statistical learning: data mining, inference, and prediction*. New York: Springer, pp. 9–40.
- He, K., Zhang, X., Ren, S. & Sun, J. (2016) Deep residual learning for image recognition. *arXiv*, 1512.03385. [Preprint].
- Heffner, E.L., Sorrells, M.E. & Jannink, J.-L. (2009) Genomic selection for crop improvement. *Crop Science*, 49, 1–12.
- Hosmer, D.W., Jr., Lemeshow, S. & Sturdivant, R.X. (2013) Applied logistic regression. In: *Wiley series in probability and statistics*, Vol. 398. Hoboken, NJ: John Wiley & Sons.
- Hu, T., Darabos, C. & Urbanowicz, R. (2020) Editorial: machine learning in genome-wide association studies. *Frontiers in Genetics*, 11, 593958.
- Hurgobin, B. & Edwards, D. (2017) SNP discovery using a pangenome: has the single reference approach become obsolete? *Biology*, 6, 21.
- Jiao, G., Shentu, X., Zhu, X., Song, W., Song, Y. & Yang, K. (2022) Utility of deep learning algorithms in initial flowering period prediction models. *Agriculture*, 12, 2161.
- Jordan, M.I. & Mitchell, T.M. (2015) Machine learning: trends, perspectives, and prospects. *Science*, 349, 255–260.
- Kattenborn, T., Leitloff, J., Schiefer, F. & Hinz, S. (2021) Review on convolutional neural networks (CNN) in vegetation remote sensing. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173, 24–49.
- Khaki, S., Wang, L. & Archontoulis, S.V. (2020) A CNN-RNN framework for crop yield prediction. *Frontiers in Plant Science*, 10, 1750.
- Kim, Y., Roh, J.-H. & Kim, H.Y. (2017) Early forecasting of rice blast disease using long short-term memory recurrent neural networks. *Sustainability*, 10, 34.
- Kole, R.K., Banerjee, H. & Bhattacharyya, A. (2001) Monitoring of market fish samples for endosulfan and hexachlorocyclohexane residues in and around calcutta. *Bulletin of Environmental Contamination and Toxicology*, 67, 554–559.
- Kushwaha, S.K., Chauhan, P., Hedlund, K. & Ahrén, D. (2015) NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction. *Bioinformatics*, 32, 1223–1225.
- Lamichhane, J.R., Dachbrodt-Saaydeh, S., Kudsk, P. & Messéan, A. (2016) Toward a reduced reliance on conventional pesticides in European agriculture. *Plant Disease*, 100, 10–24.
- Loh, P.-R., Danecek, P., Palamara, P.F., Fuchsberger, C., Reshef, A., Y., K Finucane, H. et al. (2016) Reference-based phasing using

- the haplotype reference consortium panel. *Nature Genetics*, 48, 1443–1448.
- Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *arXiv*, 1705.07874 [preprint].
- Lynch, M. & Walsh, B. (1998) *Genetics and analysis of quantitative traits*. Sunderland, MA: Sinauer.
- Madhu, G. & Rajinikanth, T.V. (2012) A novel index measure imputation algorithm for missing data values: a machine learning approach. In: *2012 IEEE international conference on computational intelligence and computing research, Coimbatore*. India: IEEE (Institute of Electrical and Electronics Engineers), pp. 1–7.
- Maharana, K., Mondal, S. & Nemade, B. (2022) A review: data pre-processing and data augmentation techniques. *Global Transitions Proceedings*, 3, 91–99.
- Maldonado, C., Mora-Poblete, F., Contreras-Soto, R.I., Ahmar, S., Chen, J.-T., do Amaral Júnior, A.T. et al. (2020) Genome-wide prediction of complex traits in two outcrossing plant species through deep learning and Bayesian regularized neural network. *Frontiers in Plant Science*, 11, 593897.
- Mohebbi, B., Tahmassebi, A., Meyer-Baese, A., Gandomi, A.H., Samui, P., Tien Bui, D. et al. (2020) Probabilistic neural networks: a brief overview of theory, implementation, and application. In: Samui, P., Bui, D.T., Chakraborty, S. & Deo, R.C. (Eds.) *Handbook of probabilistic models*. Kidlington: Butterworth-Heinemann, pp. 347–367.
- Montesinos-López, A., Rivera, C., Pinto, F., Piñera, F., Gonzalez, D., Reynolds, M. et al. (2023) Multimodal deep learning methods enhance genomic prediction of wheat breeding. *G3: Genes, Genomes, Genetics*, 13, jkad045.
- Montesinos-Lopez, O.A., Montesinos-Lopez, J.C., Salazar, E., Barron, J.A., Montesinos-Lopez, A., Buenrostro-Mariscal, R. et al. (2021) Application of a Poisson deep neural network model for the prediction of count data in genome-based prediction. *The Plant Genome*, 14, e20118.
- Montesinos-López, O.A., Montesinos-López, J.C., Singh, P., Lozano-Ramirez, N., Barrón-López, A., Montesinos-López, A. et al. (2020) A multivariate Poisson deep learning model for genomic prediction of count data. *G3: Genes, Genomes, Genetics*, 10, 4177–4190.
- Morales, A. & Villalobos, F.J. (2023) Using machine learning for crop yield prediction in the past or the future. *Frontiers in Plant Science*, 14, 1128388.
- Murtagh, F. (1991) Multilayer perceptrons for classification and regression. *Neurocomputing*, 2, 183–197.
- Nicholls, H.L., John, C.R., Watson, D.S., Munroe, P.B., Barnes, M.R. & Cabrera, C.P. (2020) Reaching the end-game for GWAS: machine learning approaches for the prioritization of complex disease loci. *Frontiers in Genetics*, 11, 350.
- Ornella, L., Pérez, P., Tapia, E., González-Camacho, J.M., Burgueño, J., Zhang, X. et al. (2014) Genomic-enabled prediction with classification algorithms. *Heredity*, 112, 616–626.
- Ornella, L., Singh, S., Perez, P., Burgueño, J., Singh, R., Tapia, E. et al. (2012) Genomic prediction of genetic values for resistance to wheat rusts. *The Plant Genome*, 5, 17.
- Pal, T., Jaiswal, V. & Chauhan, R.S. (2016) DRPPP: A machine learning based tool for prediction of disease resistance proteins in plants. *Computers in Biology and Medicine*, 78, 42–48.
- Pearce, T., Leibfried, F. & Brintrup, A. (2020) Uncertainty in neural networks: approximately Bayesian ensembling. *Proceedings of Machine Learning Research*, 108, 234–244.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2018) Scikit-learn: machine learning in python. *arXiv*, 1201.0490. [preprint].
- Pérez, P. & de los Campos, G. (2014) Genome-wide regression and prediction with the BGLR statistical package. *Genetics*, 198, 483–495.
- Pimenta, R.J.G., Aono, A.H., Burbano, R.C.V., Coutinho, A.E., da Silva, C.C., dos Anjos, I.A. et al. (2021) Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance. *Scientific Reports*, 11, 15730.
- Pimenta, R.J.G., Aono, A.H., Burbano, R.C.V., da Silva, M.F., Anjos, I.A.D., Landell, M.G.D.A. et al. (2023) Multiomic investigation of sugarcane mosaic virus resistance in sugarcane. *The Crop Journal*, 11, 1805–1815.
- Platias, C. & Petasis, G. (2020) A comparison of machine learning methods for data imputation. In: *11th Hellenic conference on artificial intelligence*. New York, NY: Association for Computer Machinery, pp. 150–159.
- Poland, J. & Rutkoski, J. (2016) Advances and challenges in genomic selection for disease resistance. *Annual Review of Phytopathology*, 54, 79–98.
- Pudjihartono, N., Fadason, T., Kempa-Liehr, A.W. & O'Sullivan, J.M. (2022) A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, 2, 927312.
- Quinlan, J.R. (1986) Induction of decision trees. *Machine Learning*, 1, 81–106.
- Rasmussen, C.E. & Williams, C.K.I. (2006) *Gaussian processes for machine learning*. Cambridge, MA: MIT Press.
- Ribeiro, M.T., Singh, S. & Guestrin, C. (2016) “Why should I trust you?” Explaining the predictions of any classifier. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Ristaino, J.B., Anderson, P.K., Bebber, D.P., Brauman, K.A., Cunniffe, N.J., Fedoroff, N.V. et al. (2021) The persistent threat of emerging plant disease pandemics to global food security. *Proceedings of the National Academy of Sciences of the United States of America*, 118, e2022239118.
- Rodrigo, H. & Tsokos, C. (2020) Bayesian modelling of nonlinear Poisson regression with artificial neural networks. *Journal of Applied Statistics*, 47, 757–774.
- Rokach, L. (2010) Ensemble-based classifiers. *Artificial Intelligence Review*, 33, 1–39.
- Rosado, R.D.S., Cruz, C.D., Barili, L.D., de Souza Carneiro, J.E., Carneiro, P.C.S., Carneiro, V.Q. et al. (2020) Artificial neural networks in the prediction of genetic merit to flowering traits in bean cultivars. *Agriculture*, 10, 638.
- Rubinsteyn, A. & Feldman, S. (2016) *Fancyimpute: an imputation library for python*. Available from: <https://github.com/iskandr/fancyimpute> [Accessed 14th August 2024]
- Sandhu, K.S., Lozada, D.N., Zhang, Z., Pumphrey, M.O. & Carter, A.H. (2021) Deep learning for predicting complex traits in spring wheat breeding program. *Frontiers in Plant Science*, 11, 613325.
- Shenge, K.C., Mabagala, R.B., Mortensen, C.N. & Wydra, K. (2014) Resistance of *Xanthomonas campestris* pv. *vesicatoria* isolates from Tanzania to copper and implications for bacterial spot management. *African Journal of Microbiology Research*, 8, 2881–2885.
- Simón, D., Borsani, O. & Filippi, C.V. (2022) RFPDR: a random forest approach for plant disease resistance protein prediction. *PeerJ*, 10, e11683.
- Sousa, I.C.D., Nascimento, M., Silva, G.N., Nascimento, A.C.C., Cruz, C.D., Silva, F.F.E. et al. (2021) Genomic prediction of leaf rust resistance to Arabica coffee using machine learning algorithms. *Scientia Agricola*, 78, e20200021.
- Sperschneider, J. (2020) Machine learning in plant-pathogen interactions: empowering biological predictions from field scale to genome scale. *New Phytologist*, 228, 35–41.
- St. Clair, D.A. (2010) Quantitative disease resistance and quantitative resistance loci in breeding. *Annual Review of Phytopathology*, 48, 247–268.
- Thabtah, F., Hammoud, S., Kamalov, F. & Gonsalves, A. (2020) Data imbalance in classification: experimental evaluation. *Information Sciences*, 513, 429–441.

- Tianqi, C. & Carlos, G. (2016) XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, California, USA*, pp. 785–794 <https://doi.org/10.1145/2939672.2939785>
- Toojinda, T., Broers, L.H., Chen, X.M., Hayes, P.M., Kleinhofs, A., Korte, J. et al. (2000) Mapping quantitative and qualitative disease resistance genes in a doubled haploid population of barley (*Hordeum vulgare*). *Theoretical and Applied Genetics*, 101, 580–589.
- Upadhyaya, S.R., Bayer, P.E., Tay Fernandez, C.G., Petereit, J., Batley, J., Bennamoun, M. et al. (2022) Evaluating plant gene models using machine learning. *Plants*, 11, 1619.
- van Klompenburg, T., Kassahun, A. & Catal, C. (2020) Crop yield prediction using machine learning: a systematic literature review. *Computers and Electronics in Agriculture*, 177, 105709.
- Wang, K., Abid, M.A., Rasheed, A., Crossa, J., Hearne, S. & Li, H. (2023) DNNGP, a deep neural network-based method for genomic prediction using multi-omics data in plants. *Molecular Plant*, 16, 279–293.
- Wang, X.-Y., Li, W.-F., Huang, Y.-K., Shan, H.-L., Zhang, R.-Y., Li, J. et al. (2019) Developing genetically segregating populations for localization of novel sugarcane brown rust resistance genes. *Euphytica*, 215, 159.
- Wang, Y., Wang, P., Guo, Y., Huang, S., Chen, Y. & Xu, L. (2021) prPred: a predictor to identify plant resistance proteins by incorporating k-spaced amino acid (group) pairs. *Frontiers in Bioengineering and Biotechnology*, 8, 645520.
- Wang, Y., Zhang, Z., Feng, L., Ma, Y. & Du, Q. (2021) A new attention-based CNN approach for crop mapping using time series Sentinel-2 images. *Computers and Electronics in Agriculture*, 184, 106090.
- Warburton, M.L., Woolfolk, S.W., Smith, J.S., Hawkins, L.K., Castano-Duque, L., Lebar, M.D. et al. (2023) Genes and genetic mechanisms contributing to fall armyworm resistance in maize. *The Plant Genome*, 16, e20311.
- Xia, J., Hu, X., Shi, F., Niu, X. & Zhang, S. (2009) Prediction of disease-resistant gene by using artificial neural network. In: *2009 international conference on research challenges in computer science*. Shanghai, China: IEEE (Institute of Electrical and Electronics Engineers), pp. 81–84.
- Xiao, Q., Li, W., Chen, P. & Wang, B. (2018) Prediction of crop pests and diseases in cotton by long short term memory network. In: Huang, D.S., Jo, K.H. & Zhang, X.L. (Eds.) *Intelligent computing theories and applications: ICIC 2018. Lecture Notes in Computer Science*. Cham: Springer, pp. 11–16.
- Xu, H., Kinfu, K.A., LeVine, W., Panda, S., Dey, J., Ainsworth, M. et al. (2021) When are deep networks really better than decision forests at small sample sizes, and how? arXiv, 2108.13637. [Preprint].
- Yang, H., Mohd Saad, N.S., Ibrahim, M.I., Bayer, P.E., Neik, T.X., Severn-Ellis, A.A. et al. (2021) Candidate *Rlm6* resistance genes against *Leptosphaeria maculans* identified through a genome-wide association study in *Brassica juncea* (L.) Czern. *Theoretical and Applied Genetics*, 134, 2035–2050.
- Young, N.D. (1996) QTL mapping and quantitative disease resistance in plants. *Annual Review of Phytopathology*, 34, 479–501.
- Zhang, F., Kang, J., Long, R., Li, M., Sun, Y., He, F. et al. (2023) Application of machine learning to explore the genomic prediction accuracy of fall dormancy in autotetraploid alfalfa. *Horticulture Research*, 10, uhac225.
- Zhang, X.D. (2020) *A matrix algebra approach to artificial intelligence*. Singapore: Springer Nature.
- Zhao, Y., Mette, M.F., Gowda, M., Longin, C.F.H. & Reif, J.C. (2014) Bridging the gap between marker-assisted and genomic selection of heading time and plant height in hybrid wheat. *Heredity*, 112, 638–645.

**How to cite this article:** Upadhyaya, S.R., Danilevicz, M.F., Dolatabadian, A., Neik, T.X., Zhang, F., Al-Mamun, H.A. et al. (2024) Genomics-based plant disease resistance prediction using machine learning. *Plant Pathology*, 00, 1–12. Available from: <https://doi.org/10.1111/ppa.13988>