Received: 15 September 2023 | Accepted: 26 February 2024

DOI: 10.1002/aps3.11603

APPLICATION ARTICLE

Applications in Plant Sciences

A machine learning algorithm for the automatic classification of *Phytophthora infestans* genotypes into clonal lineages

Camilo Patarroyo[1,2] | **Stéphane Dupas**[2] | **Silvia Restrepo**[3]

[1]Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia

[2]Université Paris-Saclay, CNRS, IRD, UMR Évolution, Génomes, Comportement et Écologie, Gif-sur-Yvette 91198, France

[3]Department of Food and Chemical Engineering, Universidad de los Andes, Bogotá, Colombia

Correspondence

Camilo Patarroyo, Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia.
Email: c.patarroyov@gmail.com

Abstract

Premise: The prompt categorization of *Phytophthora infestans* isolates into described clonal lineages is a key tool for the management of its associated disease, potato late blight. New isolates of this pathogen are currently classified by comparing their microsatellite genotypes with characterized clonal lineages, but an automated classification tool would greatly improve this process. Here, we developed a flexible machine learning–based classifier for *P. infestans* genotypes.

Methods: The performance of different machine learning algorithms in classifying *P. infestans* genotypes into its clonal lineages was preliminarily evaluated with decreasing amounts of training data. The four best algorithms were then evaluated using all collected genotypes.

Results: mlpML, cforest, nnet, and AdaBag performed best in the preliminary test, correctly classifying almost 100% of the genotypes. AdaBag performed significantly better than the others when tested using the complete data set (Tukey HSD $P < 0.001$). This algorithm was then implemented in a web application for the automated classification of *P. infestans* genotypes, which is freely available at https://github.com/cpatarroyo/genotypeclas.

Discussion: We developed a gradient boosting–based tool to automatically classify *P. infestans* genotypes into its clonal lineages. This could become a valuable resource for the prompt identification of clonal lineages spreading into new regions.

KEYWORDS

genotyping, machine learning, microsatellites, *Phytophthora infestans*, potato late blight

Phytophthora infestans ranks among the world's most economically impactful plant pathogens, causing global losses valued at an estimated annual cost of approximately 5 billion EUR (5.4 billion USD) (EuroBlight; https://agro.au.dk/forskning/internationale-platforme/euroblight/). This oomycete is the causal agent of potato late blight, which is one of the biggest threats to global food security (Fry, 2008) and has been found in almost all potato (*Solanum tuberosum* L.)-producing countries (Goodwin et al., 1994; Martin et al., 2019). These globally distributed populations are in constant flux, and changes in their composition have important implications for disease management.

Epidemics caused by *P. infestans* are mainly attributed to its rapid asexual reproduction cycles. Once in the host, this pathogen causes lesions on the leaves, producing hundreds of thousands of sporangia (Nowicki et al., 2012; Fry et al., 2015). These sporangia are aerially dispersed and can germinate directly on the host tissue or indirectly in water, producing motile infective zoospores (Judelson and Blanco, 2005; Nowicki et al., 2012; Fry et al., 2015; Whisson et al., 2016). The rapid proliferation through asexual reproduction gives rise to clonal lineages, descendants of a single recombination event, that vary only through mutation (Fry et al., 2015; Fry, 2020). Individuals that belong to a clonal lineage therefore share many phenotypic traits, including some determinants of disease management, such as fungicide resistance (Kato et al., 1997; Danies et al., 2013; Saville et al., 2015; Puidet et al., 2023), response to environmental variables (Mizubuti and Fry, 1998;

Maziero et al., 2009), aggressiveness (Njoroge et al., 2019; Puidet et al., 2022), or host range (Danies et al., 2013; Njoroge et al., 2016), to name a few. It is also important to note that some of these traits are also modulated by interactions with additional plant stressors, such as heavy metals (Arasimowicz-Jelonek et al., 2014) or coinfection with other pathogens (Kalra et al., 1989), and these environmental cofactors should be considered when phenotyping plants infected with *P. infestans*. Data sets such as the Stress Combinations and their Interactions in Plants Database (SCIPdb; Priya et al., 2023) are a valuable resource for predicting the responses of plants to specific clonal lineages of *P. infestans*, particularly when used in combination with the physiological data mentioned above.

The classification of *P. infestans* isolates into the clonal lineages is the main strategy to monitor this pathogen on a large scale (Pule et al., 2013; Njoroge et al., 2016, 2019; Nnadi et al., 2019; Guha Roy et al., 2021; Mihretu et al., 2021; Saville et al., 2021; Puidet et al., 2022, 2023). The global standard for characterizing *P. infestans* isolates involves genotyping the pathogen by amplifying 12 microsatellite loci and analyzing the sizes of the amplicons. Subsequently, each newly determined genotype is categorized by determining its closest lineage based on genetic distance (Bruvo et al., 2004; Li et al., 2013). This is still the standard method used to classify *P. infestans* populations worldwide, despite the development of new molecular markers for population genetics (Danies et al., 2013; Pule et al., 2013; Njoroge et al., 2016, 2019; Chaves et al., 2018; Dey et al., 2018; Alor et al., 2019; Martin et al., 2019; Nnadi et al., 2019; Dangi et al., 2021; Guha Roy et al., 2021; Mihretu et al., 2021; Saville et al., 2021; Puidet et al., 2022, 2023). The standardized nature of the markers used to classify these isolates into clonal lineages has allowed the monitoring of the dynamics of this pathogen on a global scale, which is performed by four international organizations: EuroBlight (https://agro.au.dk/forskning/internationale-platforme/euroblight/) in Europe, AsiaBlight (https://www.asiablight.org/) in Asia, USABlight (https://usablight.org/) in North America, and Tizón Latino (https://tizonlatino.github.io/) in Central and South America.

There are two ways to classify a multilocus genotype into a lineage. One approach is to build a dendrogram or a minimal spanning network using the unknown samples and genotypes previously classified into known clonal lineages, after which the unclassified samples can be assigned to the closest known clonal lineage (Li et al., 2013; Guha Roy et al., 2021). Alternatively, this approach can be automated (Tabima et al., 2016), as implemented in SSR Matcher (https://strain-classifier.plant-aid.org/). This second method builds a minimal spanning network with classified genotypes and new genotypes using the Bruvo genetic distance (Bruvo et al., 2004), then classifies the new genotypes into their closest clonal lineage (Tabima et al., 2016). These approaches have two important limitations: (1) they use microsatellite genotypes as their only source of information, and (2) they cannot provide detailed probabilistic information about the chance of a genotype belonging to a specific lineage. This probabilistic information would be of particular importance when specific genotypes cannot be placed into a lineage with complete certainty. Other methods, such as Bayesian phylogenetic trees (PhyML), can provide information about the probability of an unknown genotype belonging to a clonal lineage (Guindon et al., 2010); however, these require sequence information from both the unclassified and previously classified isolates, which is not typically available.

The current automatic classification algorithm proposed by Tabima et al. (2016) functionally corresponds to a k-nearest neighbors algorithm, where $k = 1$ (Kramer, 2013). Its information source is limited only to microsatellite data. The logical next step would be to develop a more general classification algorithm to expand on this idea. Although the use of machine learning (ML) algorithms has been proposed for genotypic classification in agricultural applications, these have focused mostly on the classification of plant cultivars/genotypes (Bishnoi et al., 2022), the prediction of plant phenotypes based on their genotypes (Danilevicz et al., 2022), or the prediction of pathogen phenotypes based on genomic information (te Molder et al., 2021). Most ML applications regarding *P. infestans* have been devoted to the automation of the early detection of late blight (Duarte-Carvajalino et al., 2018; Gao et al., 2021; Kool and Evenhuis, 2023; Kumar et al., 2023), not the genotypic classification of the pathogen.

In this study, we propose a ML classifier that could overcome both limitations of the current *P. infestans* classification approaches: the exclusive use of microsatellite data and the inability to report the probability of belonging to the predicted clonal lineage. The algorithm presented in this work was implemented using microsatellite information but can be easily expanded to use other genotype data, such as single-nucleotide polymorphisms (SNPs) (Schiavo et al., 2020), mitochondrial haplotypes, or even longer sequences. This is possible because these types of sequence would be encoded as categorical variables, as is the case for microsatellite data (Alkharusi, 2012; Hancock and Khoshgoftaar, 2020; Valdez-Valenzuela et al., 2021). Seven ML classification algorithms were tested for their ability to calculate the probability of each unknown element belonging to each lineage. The performance of ML algorithms is rather robust to certain data variability (Jordan and Mitchell, 2015; Sharma et al., 2021), such as that observed among genotypes of clonal lineages of *P. infestans* (Wang et al., 2017; Chaves et al., 2018; Dey et al., 2018; Fry, 2020; Lindqvist-Kreuze et al., 2020; Guha Roy et al., 2021). The best-performing algorithm was implemented in an automated genotype classifier.

METHODS

Genotype data

A data set of 1392 genotypes with 566 unique multilocus genotypes characterized by the 12 standard microsatellite loci used to describe *P. infestans* isolates (Li et al., 2013) was analyzed in this study. As all previously published genotypes were characterized using the same standard microsatellite set, they are all comparable. The genotypes composing this data set were previously classified into 23 clonal lineages. These were isolated and characterized in Colombia (Chaves et al., 2018, 2020), Peru (Lindqvist-Kreuze et al., 2020), the United States (Wang et al., 2017), and India and Europe (Dey et al., 2018).

Machine learning algorithms

As there is no consensus regarding the most effective algorithm for genotype classification (Zhao et al., 2016; Amaral et al., 2022; Bishnoi et al., 2022), seven ML classification algorithms encompassing various approaches were tested: two gradient boosting (AdaBag version 5.0 and bsttree version 0.3-24) (Hastie et al., 2009; Alfaro et al., 2013); two random forests (cforest version 1.3-14 and ORFpls version 0.3) (Breiman, 2001; Menze et al., 2011); a Bayes generalized linear model (bayesglm version 1.13-1) (Gelman et al., 2008); and two neural networks, a single hidden layer perceptron (nnet version 7.3-19) (Venables and Ripley, 2002) and a multi-layer perceptron (mlpML version 0.4-17) (Zell et al., 2011). This approach has previously been used to identify the best-suited ML algorithm for a task when there is no clear consensus (Bishnoi et al., 2022).

The general pipeline for testing all the ML models began with a data split. In this step, the genotype data set is divided into two subsets: one used to train the ML models and another used to test the performance of the trained models (as described below). The accuracy of the predicted classification is evaluated by comparing the known lineage of each genotype with the ML model's prediction, which is scored as described below (Figure 1). This general pipeline has previously been used to test the performance of ML algorithms for genotype classification (Bishnoi et al., 2022). The training of the different algorithms and the calculation of their performance metrics were performed using the caret (version 6.0-94) R package (R version 4.3.1) (Kuhn, 2008; R Core Team, 2023).

Data preparation

To prepare the data for training the models, genotype data were reorganized into a sparse matrix where each combination of locus and alleles present becomes a variable, and the value for each variable is either present (1) or absent (0) (Alkharusi, 2012; Hancock and Khoshgoftaar, 2020). This transformation was performed for three reasons: (1) it is required for including categorical values in ML models (Alkharusi, 2012; Hancock and Khoshgoftaar, 2020); (2) it is a more memory-efficient way of storing data (Cerda et al., 2018); and (3) it does not imply an order relation between the variables, as is the case for other categorical encoding procedures such as ordinal methods (Potdar et al., 2017). This sparse training data matrix was the input for training all ML models tested. The sparse matrix was the Tab slot of the Genind object imported using the *read.genalex* function of the poppr R package (version 2.9.4) (Kamvar et al., 2014).

Model training and testing

Because model training and testing are computationally intensive processes, the algorithm testing was divided into two steps. A preliminary evaluation was performed using a data set of 76 genotypes from the most represented clonal lineages (EC-1, PE-3, and EU_13), with all lineages represented in similar proportions. Both the classification accuracy and the robustness of the method with decreasing amounts of training data were assessed. The four best-performing algorithms were selected and evaluated for their classification accuracy on the complete genotype data set.

In the preliminary and final evaluations of the ML algorithms, the genotype data set was split into two parts, one used to train the algorithm (training set) and the other to test its performance (testing set), as is commonly done for this type of analysis (Bishnoi et al., 2022). Five different data splits were tested (training/test): 80%/20%, 50%/50%, 20%/80%, 10%/90% and 5%/95%. Cohen's kappa of the genotype classification in the testing set was estimated to correct for the probability of correctly classifying a genotype into a clonal lineage by chance (Warrens, 2011; Grandini et al., 2020). Cohen's kappa value oscillates between 0 and 1, where 1 indicates perfect agreement and 0 indicates a complete lack of agreement (Kuhn, 2008; Warrens, 2011; Grandini et al., 2020). This metric's correction is important in this case because different numbers of genotypes belong to each clonal lineage. For each test, 20 repetitions were run, and Cohen's kappa was calculated for each.

Testing different data splits for training and testing the models was done to determine the robustness of the classification made by each of the ML algorithms. For the preliminary evaluation, the five data splits resulted in 182 (80%), 114 (50%), 46 (20%), 23 (10%), or 11 (5%) of the 228 genotypes used for training and the remainder used for testing.

The ML algorithms that performed best along the different proportions of training data were selected for testing using the complete data set. For this final evaluation, the models were trained with 80% (1114) of the 1392 genotypes and tested with the remaining 278 (20%). For both the preliminary and complete data set tests, the classification process was repeated 20 times with randomly selected genotypes for each run, and Cohen's kappa was calculated for each run.

Performance comparison

An ANOVA test was used to compare the mean Cohen's kappa performance for the different ML algorithms when using the whole genotype data set. Tukey's honest

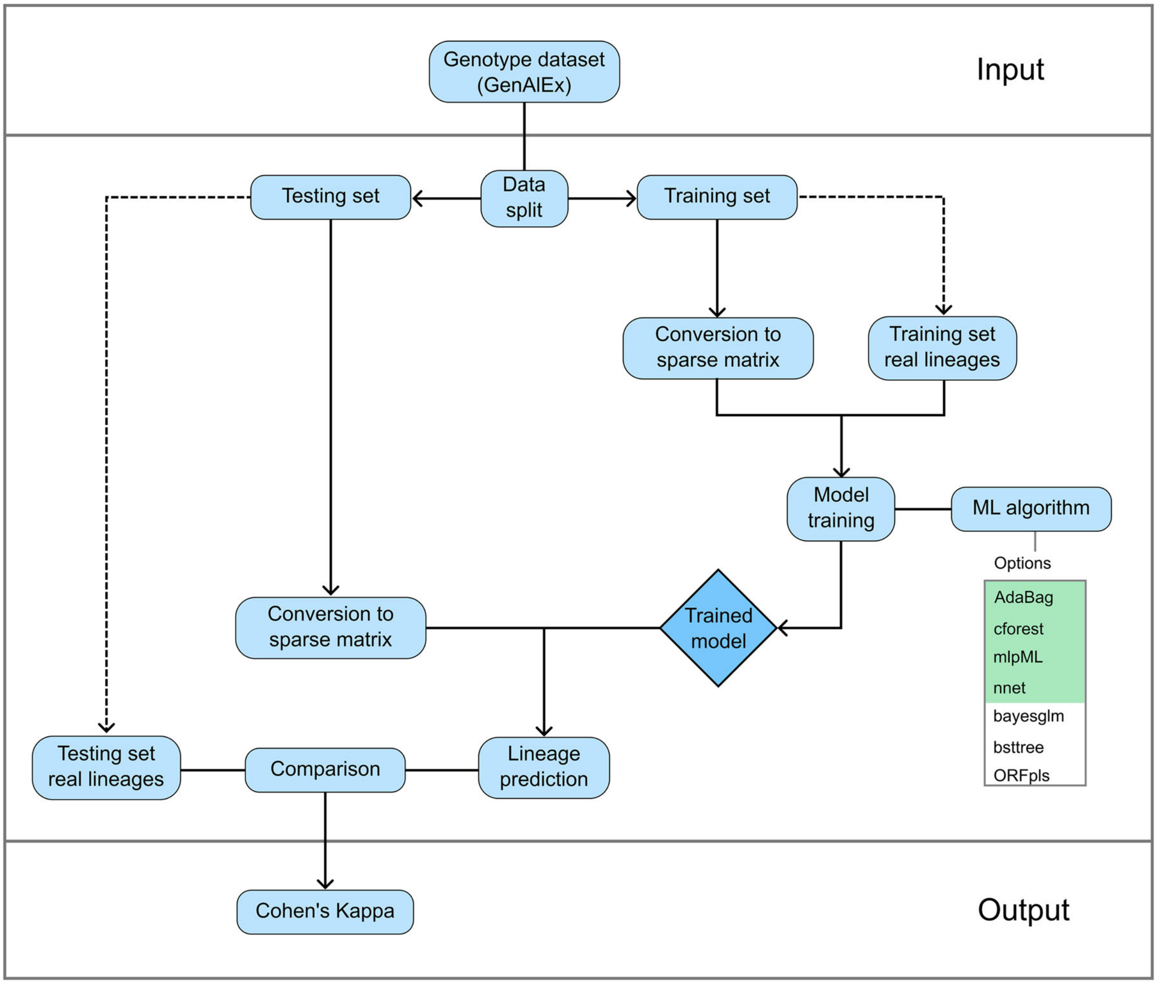


FIGURE 1 Pipeline used to test the classification models. A GenAlEx file containing both the microsatellite genotype information and the clonal lineages to which each isolate belongs is used as the input. This set was then split into a training and a testing set. The genotype data for both the training and testing sets were coded as sparse matrices. The sparse matrix and real lineages from the training set were then used to train the corresponding classification model. This training was performed in each run using one of the algorithms shown. Once the model was trained, the sparse matrix produced with the genotypes in the testing set was used to predict the lineages for the isolates in this set. These predicted lineages were compared with the real lineages to which these isolates belong, and Cohen's kappa was calculated as the output. Dotted lines represent information obtained from the genotype sets without any modification. The algorithms highlighted in green were the best performing in the preliminary test, which were subsequently tested with the complete data set. The testing process was done 20 times for each algorithm for each data split for the preliminary data set test and 20 times for the complete data set tests. ML, machine learning.

significant difference test was used to identify which of the paired differences between methods were significant. Both tests used the statistical software R version 4.2.3 (R Core Team, 2023).

RESULTS

Algorithm performance

The ML algorithms could be divided into two groups based on their performance. The first group, consisting of the bayesglm, bsttree, and ORFpls algorithms, had an average kappa value of around 0.5 for most training data proportions (Figure 2, Table 1). The only exception was the bsttree algorithm, which had a Cohen's kappa value of 0 when the proportion of data used for training dropped to 5%.

The second group, comprising AdaBag, cforest, nnet, and mlpML, had kappa values of around 1 for the 80% training data proportion and lower values for smaller training data percentages, depending on the algorithm (Figure 2, Table 1). When 5% of the genotypes were used for training, the kappa value remained above 0.99 for mlpML but decreased to around 0.86 and 0.89 for nnet and AdaBag, respectively, and dropped to 0 for cforest.

The algorithms in the second group were evaluated with the complete genotype data set (Table 2). Both nnet and cforest performed similarly ($P = 0.9996$; Table 3) in classifying the complete genotype data set, with Cohen's kappa values of around 0.61 (Figure 3, Table 2). AdaBag and mlpML performed significantly better than nnet and cforest

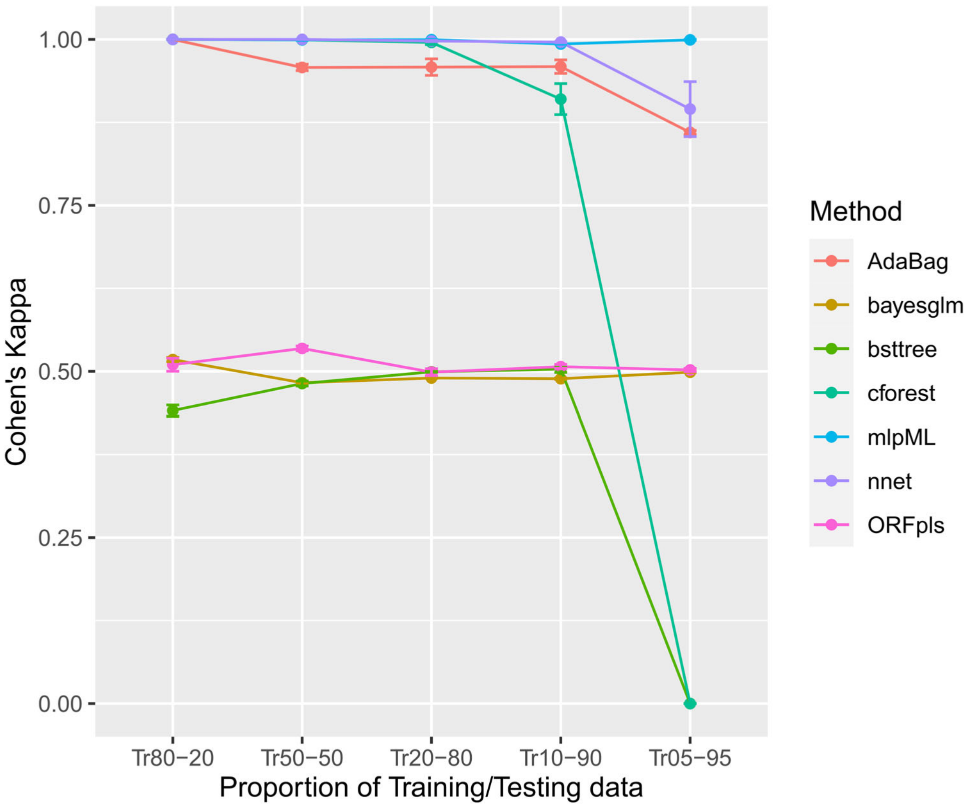


FIGURE 2 Cohen's kappa of the machine learning algorithms in classifying the genotypes in the balanced preliminary testing set. Tr80-20, 80% of the genotypes were used for training and 20% for testing; Tr50-50, 50% of the genotypes were used for training and 50% for testing; Tr20-80, 20% of the genotypes were used for training and 80% for testing; Tr10-90, 10% of the genotypes were used for training and 90% for testing; Tr05-95, 5% of the genotypes were used for training and 95% for testing. Error bars indicate the mean Cohen's kappa ± the standard error for the 20 replicates.

($P < 1e^{-10}$; Tables 2 and 3, Figure 3). AdaBag scored considerably better than mlpML ($P = 0.0004$; Tables 2 and 3), with Cohen's kappa values around 0.89 and 0.84, respectively.

Automatic classification tool

After studying the performance of the ML algorithms in classifying the *P. infestans* microsatellite genotypes into clonal lineages, both the AdaBag and the mlpML methods were implemented in a web app using the shiny R package (Chang et al., 2022). This web app takes as input a GenAlEx file (Peakall and Smouse, 2012) containing the genotypes of the samples to be classified and outputs a table with the clonal lineage predicted for each one (Figure 4). In addition, the built web app allows users to download a second table with the calculated probability of each genotype belonging to each one of the clonal lineages. All the scripts corresponding to this web app are available at https://github.com/cpatarroyo/genotypeclas (see Data Availability Statement).

The proposed workflow uses the AdaBag algorithm (Alfaro et al., 2013) to train the model using previously classified *P. infestans* genotypes. The trained model can then be used to classify all newly genotyped isolates into their corresponding clonal lineages through a user interface built using the shiny R package (version 1.7.5) (Chang et al., 2022) (Figure 4). An additional advantage of the proposed workflow (shown in Figure 5) is that the most computationally intensive part, model training, would only need to be done when new expert-vetted classified microsatellite genotypes are added to the training data set. The classification of newly genotyped samples, which would be used more often, is far less computationally intensive.

DISCUSSION

Recently, diverse ML methods have been applied to the problem of genotype classification for organisms such as plants (Sant'Anna et al., 2015; Torkzaban et al., 2015; Remita et al., 2017; Amaral et al., 2022; Nicora et al., 2022) and viruses (Remita et al., 2017). Moreover, these same methods have been used to automate specific analyses in phytopathology, such as identifying the physiological responses of potato cultivars to late blight (Gold et al., 2020). These examples illustrate a growing interest in developing ML-based tools in phytopathology, which is part of the larger trend of implementing new analytic tools in precision agriculture to more efficiently manage agricultural

TABLE 1 Summary of the performance metrics for each of the machine learning algorithms for the 20 runs of each of the different data splits performed in the preliminary test. Values of 0.0000 are below $1e^{-4}$.

Method	Split	Accuracy[a]	Kappa[b]	SD kappa[c]	TestAcc[d]	NoInfAc[e]	AccPval[f]	TestKappa[g]	SD test kappa[h]
AdaBag	Tr80-20	0.9997	0.9995	0.0022	1.0000	0.3707	0.0000	1.0000	0.0000
bsttree	Tr80-20	0.6015	0.4144	0.0169	0.5957	0.4141	0.0123	0.4409	0.0369
bayesglm	Tr80-20	0.5521	0.3756	0.0007	0.7071	0.4407	0.0003	0.5177	0.0139
cforest	Tr80-20	0.9976	1.0000	0.0000	1.0000	0.3960	0.0000	1.0000	0.0000
mlpML	Tr80-20	0.9986	0.9979	0.0036	1.0000	0.4109	0.0000	1.0000	0.0000
ORFpls	Tr80-20	0.6937	0.5357	0.0131	0.6783	0.3967	0.0006	0.5100	0.0428
nnet	Tr80-20	0.9930	0.9890	0.0027	1.0000	0.4109	0.0000	1.0000	0.0000
AdaBag	Tr50-50	0.9958	0.9936	0.0142	0.9719	0.3702	0.0000	0.9578	0.0211
bsttree	Tr50-50	0.6428	0.4574	0.0249	0.6439	0.3596	0.0000	0.4817	0.0161
bayesglm	Tr50-50	0.6152	0.4403	0.0090	0.6555	0.3454	0.0000	0.4828	0.0048
cforest	Tr50-50	0.9985	0.9977	0.0054	0.9996	0.3667	0.0000	0.9993	0.0029
mlpML	Tr50-50	0.9998	0.9998	0.0006	0.9996	0.3690	0.0000	0.9993	0.0029
ORFpls	Tr50-50	0.6892	0.5312	0.0117	0.6973	0.3559	0.0000	0.5345	0.0159
nnet	Tr50-50	0.9965	0.9945	0.0055	1.0000	0.3588	0.0000	1.0000	0.0000
AdaBag	Tr20-80	0.9570	0.9346	0.0311	0.9723	0.3451	0.0000	0.9583	0.0543
bsttree	Tr20-80	0.6321	0.4581	0.0326	0.6654	0.3360	0.0000	0.4992	0.0027
bayesglm	Tr20-80	0.6759	0.4847	0.0017	0.6573	0.3424	0.0000	0.4899	0.0021
cforest	Tr20-80	0.9735	0.9611	0.0255	0.9970	0.3492	0.0000	0.9955	0.0144
mlpML	Tr20-80	1.0000	1.0000	0.0000	0.9997	0.3486	0.0000	0.9996	0.0018
ORFpls	Tr20-80	0.6754	0.5082	0.0916	0.6647	0.3528	0.0000	0.4988	0.0225
nnet	Tr20-80	0.9963	0.9945	0.0086	0.9984	0.3527	0.0000	0.9975	0.0038
AdaBag	Tr10-90	0.8869	0.8232	0.0661	0.9727	0.3401	0.0000	0.9590	0.0443
bsttree	Tr10-90	0.5631	0.3151	0.0208	0.6698	0.3414	0.0000	0.5034	0.0222
bayesglm	Tr10-90	0.7841	0.6542	0.0611	0.6579	0.3433	0.0000	0.4889	0.0051
cforest	Tr10-90	0.7635	0.6497	0.0210	0.9398	0.3420	0.0000	0.9101	0.1018
mlpML	Tr10-90	0.9811	0.9705	0.0213	0.9954	0.3459	0.0000	0.9930	0.0016
ORFpls	Tr10-90	0.5571	0.3692	0.1261	0.6732	0.3437	0.0000	0.5070	0.0145
nnet	Tr10-90	0.9826	NA	NA	0.9973	0.3493	0.0000	0.9960	0.0036
AdaBag	Tr05-95	0.8032	0.6403	0.0344	0.9065	0.3421	0.0000	0.8598	0.0132
bsttree	Tr05-95	0.5132	NA	NA	0.3276	0.3364	0.6321	0.0000	0.0000
bayesglm	Tr05-95	0.6257	0.3222	0.0195	0.6665	0.3379	0.0000	0.4986	0.0007
cforest	Tr05-95	0.1831	NA	NA	0.3333	0.3377	0.5796	0.0000	0.0000
mlpML	Tr05-95	0.9195	0.8771	0.0234	0.9995	0.3366	0.0000	0.9993	0.0030

TABLE 1 (Continued)

Method	Split	Accuracy[a]	Kappa[b]	SD kappa[c]	TestAcc[d]	NoInfAc[e]	AccPval[f]	TestKappa[g]	SD test kappa[h]
ORFpls	Tr05-95	0.5864	0.4083	0.1849	0.6680	0.3408	0.0000	0.5021	0.0102
nnet	Tr05-95	0.8692	NA	NA	0.9293	0.3417	0.0000	0.8950	0.1803

Note: NA = not available (i.e., the metrics could not be calculated); Tr80-20 = 80% of the genotypes were used for training and 20% for testing; Tr50-50 = 50% of the genotypes were used for training and 50% for testing; Tr20-80 = 20% of the genotypes were used for training and 80% for testing; Tr10-90 = 10% of the genotypes were used for training and 90% for testing; Tr05-95 = 5% of the genotypes were used for training and 95% for testing.

[a] Average classification accuracy for the cross-validation test of the training data.

[b] Average Cohen's kappa value for the classification cross-validation test of the training data.

[c] Standard deviation of the Cohen's kappa value for the cross-validation test of the training data.

[d] Average classification accuracy of the training data set.

[e] Average classification accuracy when the genotypes were classified at random in the clonal lineages (no-information accuracy).

[f] Average *P* value of the difference between the test accuracy and the no-information accuracy.

[g] Average Cohen's kappa value for the classification of the testing data.

[h] Standard deviation of the Cohen's kappa value for the classification of the testing data.

TABLE 2 Summary of performance metrics for each of the machine learning algorithms in the complete data set test using 80% of the genotypes for training and 20% for testing. Values of 0.0000 are below $1e^{-4}$.

Method	Accuracy[a]	Kappa[b]	SD kappa[c]	TestAcc[d]	NoInfAc[e]	AccPval[f]	TestKappa[g]	SD test kappa[h]
AdaBag	0.9043	0.8512	0.0126	0.9579	0.7790	0.0000	0.8879	0.0210
bsttree	0.6403	0.4340	0.0065	0.8316	0.7643	0.0059	0.5760	0.0184
bayesglm	0.6529	0.4582	0.0030	0.8281	0.7810	0.0322	0.5495	0.0048
cforest	0.8217	0.6558	0.0093	0.8837	0.7739	0.0001	0.6100	0.0460
mlpML	0.9031	0.8426	0.0131	0.9414	0.7691	0.0000	0.8393	0.0484
nnet	0.7082	0.5587	0.0168	0.8411	0.7643	0.0017	0.6088	0.0115

[a] Average classification accuracy for the cross-validation test of the training data.

[b] Average Cohen's kappa value for the classification cross-validation test of the training data.

[c] Standard deviation of the Cohen's kappa value for the cross-validation test of the training data.

[d] Average classification accuracy of the training data set.

[e] Average classification accuracy when the genotypes were classified at random in the clonal lineages (no-information accuracy).

[f] Average *P* value of the difference between the test accuracy and the no-information accuracy.

[g] Average Cohen's kappa value for the classification of the testing data.

[h] Standard deviation of the Cohen's kappa value for the classification of the testing data.

TABLE 3 Differences in the mean Cohen's kappa values among pairs of machine learning methods calculated using Tukey's honest significant difference method. The difference between means, the lower and upper values of the 95% confidence interval for the difference, and the *P* value after correcting for multiple comparisons are presented.

Method comparison	Difference	Lower	Upper	*P* adjusted
cforest–AdaBag	−0.2779	−0.3081	−0.2477	0.0000
mlpML–AdaBag	−0.0486	−0.0788	−0.0184	0.0004
nnet–AdaBag	−0.2791	−0.3093	−0.2489	0.0000
mlpML–cforest	0.2293	0.1991	0.2595	0.0000
nnet–cforest	−0.0012	−0.0314	0.0290	0.9996
nnet–mlpML	−0.2305	−0.2607	−0.2003	0.0000

activities, focused in this case on the protection against pathogens (Sharma et al., 2021). It is therefore beneficial to develop a ML-based algorithm for classifying *P. infestans* genotypes into clonal lineages.

The only other automatic approach to classifying *P. infestans* genotypes, SSR Matcher (Tabima et al., 2016), works as a specific type of *k*-nearest neighbors, a type of ML algorithm, for which $k = 1$ (Kramer, 2013). This algorithm reduces the differences in the variables between objects into a distance and classifies the new object in the same class as the *k* neighbors closest to it (with the least distance). In the case of SSR Matcher, the differences between the microsatellite markers are reduced to the Bruvo distance (Bruvo et al., 2004), and the unknown genotype is classified into the closest clonal lineage neighbor (Tabima et al., 2016). This approach is limited to genetic distances between microsatellite markers calculated by the Bruvo metric; however, this can be overcome by using other ML approaches that are not restricted to these types of markers and metrics.

Any categorical variable (such as mating type or mitochondrial haplotype) can be converted into a sparse matrix (also known as one-hot encoding) (Alkharusi, 2012; Hancock and Khoshgoftaar, 2020) and added to these ML methods without any changes. Numerical variables (such as

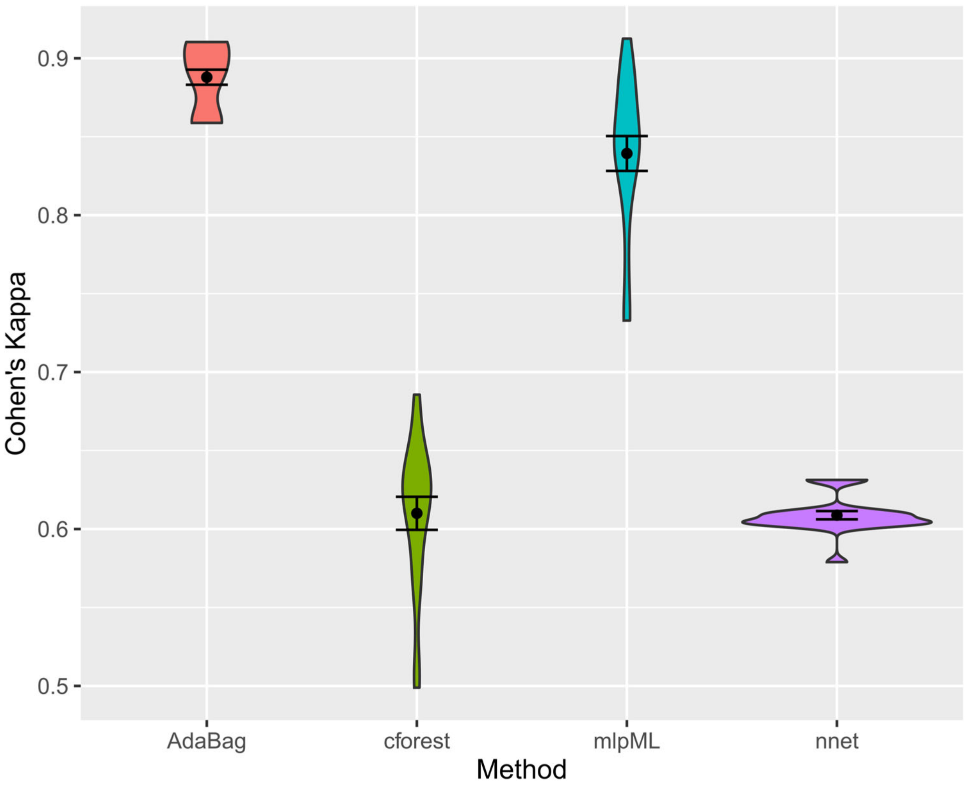


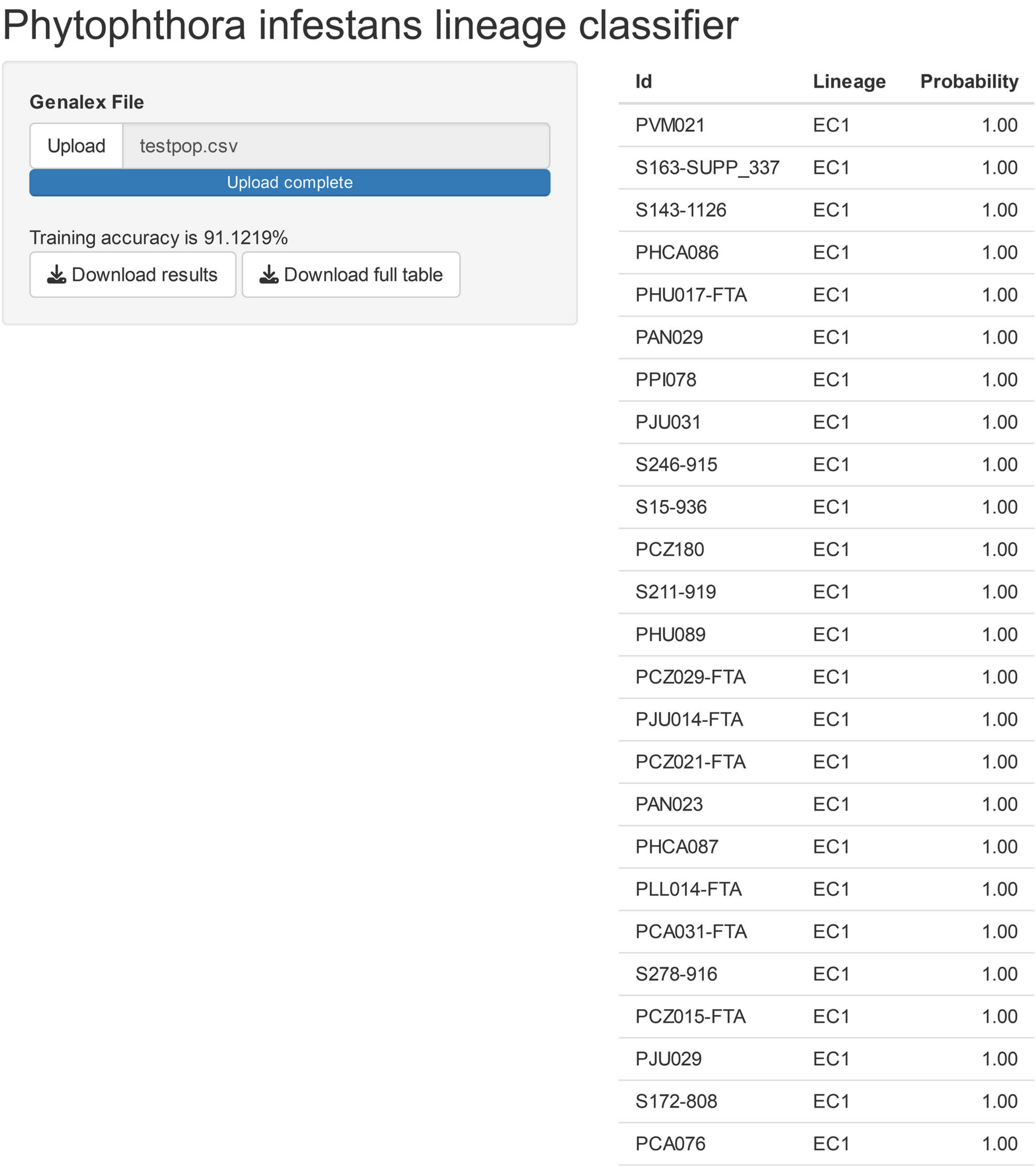
FIGURE 3 Cohen's kappa for the classification of the genotypes using AdaBag, cforest, mlpML, and nnet on the complete data set. The training set comprised 80% of the genotypes, while 20% were used for testing. The mean Cohen's kappa value and standard error for the 20 replicates for each method are represented by the dot and whiskers, respectively.

phenotypic characteristics) could also be included without modifying these ML classification algorithms. For these reasons, the ML-based classification approach presented here is significantly more flexible than SSR Matcher (Tabima et al., 2016), requiring no significant changes if the references used to genotype the *P. infestans* isolates are changed altogether from microsatellites to other molecular markers. This is particularly important in the context of the increasing numbers of molecular markers available due to the use of new sequencing technologies. Another advantage of the present approach is that the probability of each of the unknown genotypes belonging to each clonal lineage can be calculated (Kuhn, 2008). This is of particular interest to closely examine classifications that might seem incorrect or for the detection of newly formed clonal lineages.

Interestingly, the bayesglm, bsttree, and ORFpls algorithms performed consistently poorly when classifying the *P. infestans* genotypes into their corresponding clonal lineages. The classifications predicted by bsttree were no better than a random allocation when using the smallest training set. AdaBag, cforest, mlpML, and nnet performed much better in the preliminary test with a balanced training set (composed of roughly the same number of representatives from each clonal lineage); however, the progressive reduction of training information had a larger effect on AdaBag, cforest, and nnet, whereas mlpML was only slightly affected by it. This suggests that the mlpML algorithm is very robust even when training data are reduced, as long as its categories are equilibrated.

When tested on the entire data set, the performance of all four algorithms decreased, with cforest and nnet having significantly lower Cohen's kappa scores than AdaBag and mlpML. It is important to note that this data set had an additional complication: some clonal lineages were over-represented (e.g., EC-1), while others were represented by one or two genotypes (e.g., EU-8). This imbalance could be one of the main reasons for the decrease in accuracy across all methods. Despite this challenge, both AdaBag and mlpML maintained high classification accuracy. The fact that AdaBag was significantly more accurate than mlpML could indicate that this algorithm is more resilient to unbalanced training information in this case. On the other hand, mlpML is more resilient to reduced information if it is less unbalanced. These results highlight the importance of having a balanced training set for accurate classification, even if this means removing some genotypes. For automated classification, it is recommended to prioritize a balanced training set over a larger one.

Although there is no clear consensus on the best algorithm for classifying genotypes, our results are consistent with those obtained for a cotton (*Gossypium hirsutum* L.) genotype classifier (Bishnoi et al., 2022), where an algorithm based on the same principle of AdaBag (AdaBoost) showed the best performance. These findings differ from other biological classification studies that found either random



Phytophthora infestans lineage classifier

Training accuracy is 91.1219%

Id	Lineage	Probability
PVM021	EC1	1.00
S163-SUPP_337	EC1	1.00
S143-1126	EC1	1.00
PHCA086	EC1	1.00
PHU017-FTA	EC1	1.00
PAN029	EC1	1.00
PPI078	EC1	1.00
PJU031	EC1	1.00
S246-915	EC1	1.00
S15-936	EC1	1.00
PCZ180	EC1	1.00
S211-919	EC1	1.00
PHU089	EC1	1.00
PCZ029-FTA	EC1	1.00
PJU014-FTA	EC1	1.00
PCZ021-FTA	EC1	1.00
PAN023	EC1	1.00
PHCA087	EC1	1.00
PLL014-FTA	EC1	1.00
PCA031-FTA	EC1	1.00
S278-916	EC1	1.00
PCZ015-FTA	EC1	1.00
PJU029	EC1	1.00
S172-808	EC1	1.00
PCA076	EC1	1.00

FIGURE 4 User interface and output of the genotype classification tool developed using the AdaBag algorithm. Inside the box at the top left is the "Upload" button for uploading a GenAlEx file with the microsatellite information of the genotyped isolates. Below this button is a "Training accuracy" letterbox showing the cross-validation training accuracy for the training data. There are two additional buttons, "Download results" for downloading the results table displayed and "Download full table" for downloading a table with the probability of each genotype belonging to each one of the clonal lineages present in the training data set.

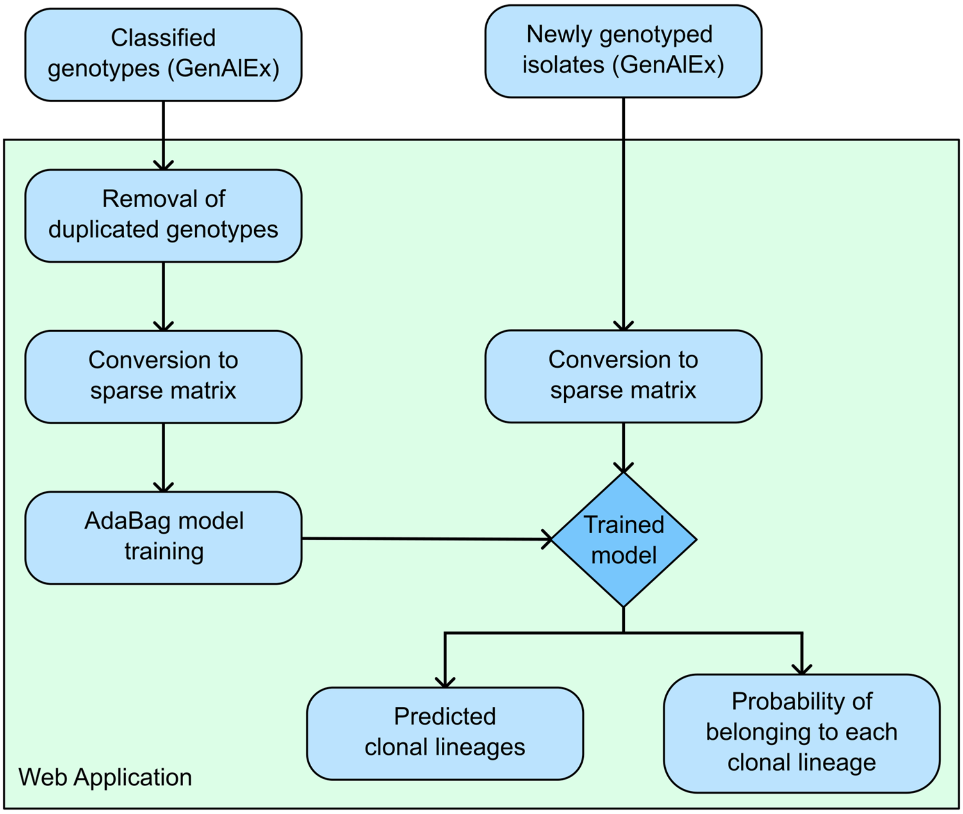

Classified genotypes (GenAlEx)
Newly genotyped isolates (GenAlEx)
Removal of duplicated genotypes
Conversion to sparse matrix
Conversion to sparse matrix
AdaBag model training
Trained model
Predicted clonal lineages
Probability of belonging to each clonal lineage
Web Application

FIGURE 5 Proposed workflow for the developed classification tool. The properly classified genotypes serve as the training data for the model, which uses the AdaBag algorithm. The trained model can then be used to classify the newly genotyped isolates. The green square contains the processes that are automated/performed by the developed web application.

forests (Schiavo et al., 2020; Borkenhagen et al., 2021), support vector machines (Athamanolap et al., 2014; Borkenhagen et al., 2021), or artificial neural networks to have the best performance (Sant'Anna et al., 2015; Borkenhagen et al., 2021; Amaral et al., 2022). This highlights the need to continue exploring different ML algorithms for addressing these biological classification problems.

The automated classification of newly genotyped *P. infestans* isolates using ML approaches is faster and more computationally efficient than the current method using SRR Matcher. Our work also highlighted potential ways to further improve the functioning of this classifier; however, some would not be practical to enact. For example, the inclusion of phenotypic characteristics as an additional source of information could improve the classification accuracy, but the determination of these physiological traits requires additional time-consuming experiments (Kato et al., 1997; Mizubuti and Fry, 1998; Maziero et al., 2009; Danies et al., 2013; Saville et al., 2015; Njoroge et al., 2016, 2019; Puidet et al., 2022, 2023), which would defeat the purpose of being a quick and efficient monitoring tool.

Other recommendations could maintain the efficiency and practicality of this tool while potentially improving its accuracy. For example, instead of one-hot encoding, different ways of encoding the categorical variables could be tested for their potential to increase the predictive performance of the models (Potdar et al., 2017; Hancock and Khoshgoftaar, 2020; Dahouda and Joe, 2021; Valdez-Valenzuela et al., 2021; Cerda and Varoquaux, 2022). Furthermore, as the results suggest, it is important to use a training data set that captures the variability of the genotypes included while maintaining a balanced representation of the lineages when considering deploying this web app for general use. This work used publicly available *P. infestans* genotype data for the model training and testing, but these data do not capture the full variability of this pathogen's genotypes. This classifier could greatly benefit from access to expertly curated classified genotype data sets, such as the ones maintained by Euro-Blight (https://agro.au.dk/forskning/internationale-platforme/euroblight/), AsiaBlight (https://www.asiablight.org/), USA-Blight (https://usablight.org/), and the Tizón Latino (https://tizonlatino.github.io/) consortia. With access to these large training data sets, two of the biggest advantages of ML models become evident. First, larger and more diverse training data sets tend to result in more accurate predictions by ML models (Shalev-Shwartz et al., 2012; Cho et al., 2015; Johnson et al., 2018; Punia et al., 2021), and second, the automation of the classification of newly genotyped isolates is faster and more computationally efficient than the current approach (Tabima et al., 2016).

As discussed above, the automatic genotype classification tool presented in this work could be refined and expanded in many ways to improve its functionality. It could also complement or be complemented by other tools in this field; for example, it can complement tools such as SSR Matcher to effectively monitor the dynamics of *P. infestans*. It could also be paired with tools such as SCIPdb (Priya et al., 2023) to improve the recommendations for growers and researchers considering the predicted clonal lineages and the interactions with other stressors in each case. This approach could be taken even further by including some of the underlying phenotypes or biological mechanisms responsible for the genesis of these lineages. Overall, the approach presented in this work represents a novel, flexible, efficient, and accurate way to automate the classification of *P. infestans* genotypes into its clonal lineages, which could prove valuable in the monitoring of this pathogen.

AUTHOR CONTRIBUTIONS

C.P. performed the analysis of the machine learning algorithms and developed the *P. infestans* classification tool. C.P., S.D., and S.R. were all responsible for the conceptualization of the study, and writing and editing the manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors are grateful for the support provided by the Institut de Recherche pour le Développement (IRD) through the Allocations de Recherche pour une Thèse au Sud (ARTS) scholarship program attributed to the Evolution, Génomes, Comportement, Ecologie (EGCE) laboratory (scholarship number 122837U, CNRS-Paris-Saclay University-IRD UMR9191), and for funding provided by the Faculty of Sciences, Universidad de los Andes, Bogotá, Colombia.

DATA AVAILABILITY STATEMENT

All the scripts used for the training, testing, and implementation of the machine learning classification models in the web app are freely available at https://github.com/cpatarroyo/genotypeclas.

OPEN RESEARCH BADGES

This article has earned an Open Materials badge for making publicly available the components of the research methodology needed to reproduce the reported procedure and analysis. All materials are available at https://github.com/cpatarroyo/genotypeclas.

ORCID

Camilo Patarroyo http://orcid.org/0000-0002-3124-1910
Stéphane Dupas http://orcid.org/0000-0001-8332-6679
Silvia Restrepo http://orcid.org/0000-0001-9016-1040

REFERENCES

Alfaro, E., M. Gáamez, and N. García. 2013. adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software* 54(2): 1–35. https://doi.org/10.18637/JSS.V054.I02

Alkharusi, H. 2012. Categorical variables in regression analysis: A comparison of dummy and effect coding. *International Journal of Education* 4: 202–210. https://doi.org/10.5296/ije.v4i2.1962

Alor, N., R. Tierno, D. E. L. Cooke, and J. I. R. de Galarreta. 2019. Characterisation of *Phytophthora infestans* isolates of potato crops from Spain. *Potato Research* 62(4): 453–463. https://doi.org/10.1007/s11540-019-9422-7

Amaral, L. de O., G. V. Miranda, B. H. P. Val, A. P. Silva, A. C. R. Moitinho, and S. H. Unêda-Trevisoli. 2022. Artificial neural network for discrimination and classification of tropical soybean genotypes of different relative maturity groups. *Frontiers in Plant Science* 13: 814046. https://doi.org/10.3389/FPLS.2022.814046

Arasimowicz-Jelonek, M., J. Floryszak-Wieczorek, K. Drzewiecka, J. Chmielowska-Bąk, D. Abramowski, and K. Izbiańska. 2014. Aluminum induces cross-resistance of potato to *Phytophthora infestans*. *Planta* 239(3): 679–694. https://doi.org/10.1007/s00425-013-2008-8

Athamanolap, P., V. Parekh, S. I. Fraley, V. Agarwal, D. J. Shin, M. A. Jacobs, T. H. Wang, and S. Yang. 2014. Trainable high resolution melt curve machine learning classifier for large-scale reliable genotyping of sequence variants. *PLoS ONE* 9(10): e109094. https://doi.org/10.1371/journal.pone.0109094

Bishnoi, S., N. Al-Ansari, M. Khan, S. Heddam, and A. Malik. 2022. Classification of cotton genotypes with mixed continuous and categorical variables: Application of machine learning models. *Sustainability (Switzerland)* 14(20): 13685. https://doi.org/10.3390/su142013685

Borkenhagen, L. K., M. W. Allen, and J. A. Runstadler. 2021. Influenza virus genotype to phenotype predictions through machine learning: A systematic review. *Emerging Microbes and Infections* 10(1): 1896–1907. https://doi.org/10.1080/22221751.2021.1978824

Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32. https://doi.org/10.1023/A:1010933404324/METRICS

Bruvo, R., N. K. Muchels, T. G. D'Souza, and H. Schulenburg. 2004. A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Molecular Ecology* 13(7): 2101–2106. https://doi.org/10.1111/j.1365-294X.2004.02209.x

Cerda, P., and G. Varoquaux. 2022. Encoding high-cardinality string categorical variables. *IEEE Transactions on Knowledge and Data Engineering* 34(3): 1164–1176. https://doi.org/10.1109/TKDE.2020.2992529

Cerda, P., G. Varoquaux, and B. Kégl. 2018. Similarity encoding for learning with dirty categorical variables. *Machine Learning* 107(8–10): 1477–1494. https://doi.org/10.1007/s10994-018-5724-2

Chang, W., J. Cheng, J. Allaire, C. Sievert, B. Scholerke, Y. Xie, J. Allen, et al. 2022. shiny: Web Application Framework for R. https://cran.r-project.org/package=shiny [accessed 3 April 2024].

Chaves, S. C., M. C. Rodríguez, M. F. Mideros, F. Lucca, C. E. Ñústez, and S. Restrepo. 2018. Determining whether geographic origin and potato genotypes shape the population structure of *Phytophthora infestans* in the Central Region of Colombia. *Phytopathology* 109(1): 145–154. https://doi.org/10.1094/PHYTO-05-18-0157-R

Chaves, S. C., N. Guayazán, M. F. Mideros, M. Parra, F. Lucca, and S. Restrepo. 2020. Two clonal species of *Phytophthora* associated to solanaceous crops coexist in Central and Southern Colombia. *Phytopathology* 110(7): 1342–1351. https://doi.org/10.1094/PHYTO-05-19-0175-R

Cho, J., K. Lee, E. Shin, G. Choy, and S. Do. 2015. How much data is needed to train a medical image deep learning system to achieve necessary high accuracy? *ArXiv* 1511.06348 [Preprint]. Published 19 November 2015 [accessed 3 April 2024]. Available from: https://doi.org/10.48550/arXiv.1511.06348.

Dahouda, M. K., and I. Joe. 2021. A deep-learned embedding technique for categorical features encoding. *IEEE Access* 9: 114381–114391. https://doi.org/10.1109/ACCESS.2021.3104357

Dangi, S., P. Wharton, A. D. Ambarwati, T. J. Santoso, Kusmana, I. Sulastrini, J. Medendorp, et al. 2021. Genotypic and phenotypic characterization of *Phytophthora infestans* populations on Java, Indonesia. *Plant Pathology* 70(1): 61–73. https://doi.org/10.1111/ppa.13269

Danies, G., I. M. Small, K. Myers, R. Childers, and W. E. Fry. 2013. Phenotypic characterization of recent clonal lineages of *Phytophthora* *infestans* in the United States. *Plant Disease* 97(7): 873–881. https://doi.org/10.1094/PDIS-07-12-0682-RE

Danilevicz, M. F., M. Gill, R. Anderson, J. Batley, M. Bennamoun, P. E. Bayer, and D. Edwards. 2022. Plant genotype to phenotype prediction using machine learning. *Frontiers in Genetics* 13: 822173. https://doi.org/10.3389/fgene.2022.822173

Dey, T., A. Saville, K. Myers, S. Tewari, D. E. L. Cooke, S. Tripathy, W. E. Fry, et al. 2018. Large sub-clonal variation in *Phytophthora infestans* from recent severe late blight epidemics in India. *Scientific Reports* 8(1): 4429. https://doi.org/10.1038/s41598-018-22192-1

Duarte-Carvajalino, J. M., D. F. Alzate, A. A. Ramirez, J. D. Santa-Sepulveda, A. E. Fajardo-Rojas, and M. Soto-Suárez. 2018. Evaluating late blight severity in potato crops using unmanned aerial vehicles and machine learning algorithms. *Remote Sensing* 10(10): 1513. https://doi.org/10.3390/rs10101513

Fry, W. E. 2008. *Phytophthora infestans*: The plant (and R gene) destroyer. *Molecular Plant Pathology* 9(3): 385–402. https://doi.org/10.1111/j.1364-3703.2007.00465.x

Fry, W. E. 2020. *Phytophthora infestans*: The itinerant invader; "late blight": the persistent disease. *Phytoparasitica* 48(1): 87–94. https://doi.org/10.1007/s12600-019-00778-3

Fry, W. E., P. R. Birch, H. S. Judelson, N. J. Grunwald, G. Danies, K. L. Everts, A. J. Gevens, et al. 2015. Five reasons to consider *Phytophthora infestans* a reemerging pathogen. *Phytopathology* 105(7): 966–981. https://doi.org/10.1094/PHYTO-01-15-0005-FI

Gao, J., J. C. Westergaard, E. H. R. Sundmark, M. Bagge, E. Liljeroth, and E. Alexandersson. 2021. Automatic late blight lesion recognition and severity quantification based on field imagery of diverse potato genotypes by deep learning. *Knowledge-Based Systems* 214: 106723. https://doi.org/10.1016/j.knosys.2020.106723

Gelman, A., A. Jakulin, M. G. Pittau, and Y. S. Su. 2008. A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics* 2(4): 1360–1383. https://doi.org/10.1214/08-AOAS191

Gold, K. M., P. A. Townsend, I. Herrmann, and A. J. Gevens. 2020. Investigating potato late blight physiological differences across potato cultivars with spectroscopy and machine learning. *Plant Science* 295: 110316. https://doi.org/10.1016/J.PLANTSCI.2019.110316

Goodwin, S. B., B. A. Cohen, and W. E. Fry. 1994. Panglobal distribution of a single clonal lineage of the Irish potato famine fungus. *Proceedings of the National Academy of Sciences, USA* 91(24): 11591–11595. https://doi.org/10.1073/pnas.91.24.11591

Grandini, M., E. Bagli, and G. Visani. 2020. Metrics for multi-class classification: An overview. *ArXiv* 2008.05756 [Preprint]. Published 13 August 2020 [accessed 3 April 2024]. Available from: https://doi.org/10.48550/arxiv.2008.05756

Guha Roy, S., T. Dey, D. E. L. Cooke, and L. R. Cooke. 2021. The dynamics of *Phytophthora infestans* populations in the major potato-growing regions of Asia – A review. *Plant Pathology* 70(5): 1015–1031. https://doi.org/10.1111/ppa.13360

Guindon, S., J. F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology* 59(3): 307–321. https://doi.org/10.1093/SYSBIO/SYQ010

Hancock, J. T., and T. M. Khoshgoftaar. 2020. Survey on categorical data for neural networks. *Journal of Big Data* 7(1): 28. https://doi.org/10.1186/s40537-020-00305-w

Hastie, T., S. Rosset, J. Zhu, and H. Zou. 2009. Multi-class AdaBoost. *Statistics and Its Interface* 2(3): 349–360. https://doi.org/10.4310/SII.2009.V2.N3.A8

Johnson, M., P. Anderson, M. Dras, and M. Steedman. 2018. Predicting accuracy on large datasets from smaller pilot data. *In* I. Gurevych and Y. Miyao [eds.], Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 450–455. Melbourne, Australia. https://doi.org/10.18653/v1/P18-2072

Jordan, M. I., and T. M. Mitchell. 2015. Machine learning: Trends, perspectives, and prospects. *Science* 349(6245): 255–260. https://doi.org/10.1126/science.aaa8415

Judelson, H. S., and F. A. Blanco. 2005. The spores of *Phytophthora*: Weapons of the plant destroyer. *Nature Reviews Microbiology* 3(1): 47–58. https://doi.org/10.1038/nrmicro1064

Kalra, A., R. K. Grover, N. Rishi, and S. M. P. Khurana. 1989. Interaction between *Phytophthora infestans* and potato viruses X and Y in potato. *Journal of Agricultural Science* 112: 33–37. https://doi.org/10.1017/S0021859600084070

Kamvar, Z. N., J. F. Tabima, and N. J. Grünwald. 2014. *Poppr*: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ* 2: e281. https://doi.org/10.7717/peerj.281

Kato, M., E. Mizubuti, S. Goodwin, and W. E. Fry. 1997. Sensitivity to protectant fungicides and pathogenic fitness of clonal lineages of *Phytophthora infestans* in the United States. *Phytopathology* 87(9): 973–978. https://doi.org/10.1094/PHYTO.1997.87.9.973

Kool, J., and A. Evenhuis. 2023. Early detection of *Phytophthora infestans* in potato plants using hyperspectral imaging, local comparison and a convolutional neural network. *Smart Agricultural Technology* 6: 100333. https://doi.org/10.1016/j.atech.2023.100333

Kramer, O. 2013. K-nearest neighbors. *In* O. Kramer [ed.], *Dimensionality Reduction with Unsupervised Nearest Neighbors*, 13–23. Springer, Berlin, Germany. https://doi.org/10.1007/978-3-642-38652-7_2

Kuhn, M. 2008. Building predictive models in R using the caret package. *Journal of Statistical Software* 28(5): 1–26. https://doi.org/10.18637/JSS.V028.I05

Kumar, P. K., I. Kumar, S. Kumar, P. Kumar, J. Harshith, and A. Dutta. 2023. Diagnosing *Phytophthora infestans* infestations on *Solanum tuberosum* leaves using machine learning classifiers. *In* 2023 International Conference on Computational Intelligence and Sustainable Engineering Solutions (CISES), 95–99. G.L. Bajaj Institute of Technology and Management, Greater Noida, India. https://doi.org/10.1109/CISES58720.2023.10183419

Li, Y., D. E. L. Cooke, E. Jacobsen, and T. van der Lee. 2013. Efficient multiplex simple sequence repeat genotyping of the oomycete plant pathogen *Phytophthora infestans*. *Journal of Microbiological Methods* 92(3): 316–322. https://doi.org/10.1016/j.mimet.2012.11.021

Lindqvist-Kreuze, H., S. Gamboa, M. Izarra, W. Pérez, M. Y. Correa, A. Astete, T. Särkinen, et al. 2020. Population structure and host range of the potato late blight pathogen *Phytophthora infestans* in Peru spanning two decades. *Plant Pathology* 69(2): 334–346. https://doi.org/10.1111/ppa.13125

Martin, F. N., Y. Zhang, D. Cooke, M. Coffey, N. Grünwald, and W. E. Fry. 2019. Insights into evolving global populations of *Phytophthora infestans* via new complementary mtDNA haplotype markers and nuclear SSRs. *PLoS ONE* 14(1): e0208606. https://doi.org/10.1371/journal.pone.0208606

Maziero, J. M. N., L. A. Maffia, and E. S. Mizubuti. 2009. Effects of temperature on events in the infection cycle of two clonal lineages of *Phytophthora infestans* causing late blight on tomato and potato in Brazil. *Plant Disease* 93(5): 459–466. https://doi.org/10.1094/PDIS-93-5-0459

Menze, B. H., B. M. Kelm, D. N. Splitthoff, U. Koethe, and F. A. Hamprecht. 2011. On oblique random forests. *In* D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis [eds.], Machine Learning and Knowledge Discovery in Databases. ECML PKDD 2011. Lecture Notes in Computer Science, volume 6912, 453–469. https://doi.org/10.1007/978-3-642-23783-6_29

Mihretu, E., M. Izarra, H. Lindqvist-Kreuze, W. Mohammod, and B. Kassa. 2021. Population structure of *Phytophthora infestans* (Mont.) de Bary in Ethiopia. *Journal of Plant Pathology* 103(3): 759–767. https://doi.org/10.1007/S42161-021-00820-6

Mizubuti, E., and W. E. Fry. 1998. Temperature effects on developmental stages of isolates from three clonal lineages of *Phytophthora infestans*. *Phytopathology* 88(8): 837–843. https://doi.org/10.1094/PHYTO.1998.88.8.837

Nicora, G., S. Zucca, I. Limongelli, R. Bellazzi, and P. Magni. 2022. A machine learning approach based on ACMG/AMP guidelines for genomic variant classification and prioritization. *Scientific Reports* 12(1): 2517. https://doi.org/10.1038/s41598-022-06547-3

Njoroge, A. W., G. Tusiime, G. A. Forbes, and J. E. Yuen. 2016. Displacement of US-1 clonal lineage by a new lineage of *Phytophthora infestans* on potato in Kenya and Uganda. *Plant Pathology* 65(4): 587–592. https://doi.org/10.1111/ppa.12451

Njoroge, A. W., B. Andersson, J. E. Yuen, and G. A. Forbes. 2019. Greater aggressiveness in the 2_A1 lineage of *Phytophthora infestans* may partially explain its rapid displacement of the US-1 lineage in east Africa. *Plant Pathology* 68(3): 566–575. https://doi.org/10.1111/ppa.12977

Nnadi, N. E., A. M. Datiri, D. B. Pam, A. C. Ngene, F. O. Okonkwo, L. Sullivan, and D. E. L. Cooke. 2019. First report of the EU_33_A2 clonal lineage of *Phytophthora infestans* causing late blight disease of potato in Nigeria. *New Disease Reports* 40(1): 20. https://doi.org/10.5197/j.2044-0588.2019.040.020

Nowicki, M., M. R. Foolad, M. Nowakowska, and E. U. Kozik. 2012. Potato and tomato late blight caused by *Phytophthora infestans*: An overview of pathology and resistance breeding. *Plant Disease* 96(1): 4–17. https://doi.org/10.1094/PDIS-05-11-0458

Peakall, R., and P. E. Smouse. 2012. GenAlEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research--an update. *Bioinformatics* 28(19): 2537–2539. https://doi.org/10.1093/bioinformatics/bts460

Potdar, K., T. Pardawala, and C. Pai. 2017. A comparative study of categorical variable encoding techniques for neural network classifiers. *International Journal of Computer Applications* 175(4): 7–9. https://doi.org/10.5120/ijca2017915495

Priya, P., M. Patil, P. Pandey, A. Singh, V. S. Babu, and M. Senthil-Kumar. 2023. Stress combinations and their interactions in plants database: A one-stop resource on combined stress responses in plants. *Plant Journal* 116(4): 1097–1117. https://doi.org/10.1111/tpj.16497

Puidet, B., R. Mabon, M. Guibert, R. Kiiker, L. Soonvald, V. H. Le, H. Eikemo, et al. 2022. Examining phenotypic traits contributing to the spread in Northern European potato crops of EU_41_A2, a new clonal lineage of *Phytophthora infestans*. *Phytopathology* 112(2): 414–421. https://doi.org/10.1094/phyto-12-20-0542-r

Puidet, B., R. Mabon, M. Guibert, R. Kiiker, K. Loit, V. H. Le, H. Eikemo, et al. 2023. Investigating phenotypic traits as potential drivers of the emergence of EU_37_A2, an invasive new lineage of *Phytophthora infestans* in Western Europe. *Plant Pathology* 72(4): 797–806. https://doi.org/10.1111/ppa.13700

Pule, B. B., J. C. Meitz, A. H. Thompson, C. C. Linde, W. E. Fry, S. D. Langenhoven, K. L. Meyers, et al. 2013. *Phytophthora infestans* populations in central, eastern and southern African countries consist of two major clonal lineages. *Plant Pathology* 62(1): 154–165. https://doi.org/10.1111/J.1365-3059.2012.02608.X

Punia, S. K., M. Kumar, T. Stephan, G. G. Deverajan, and R. Patan. 2021. Performance analysis of machine learning algorithms for big data classification: ML and AI-based algorithms for big data analysis. *International Journal of E-Health and Medical Communications* 12(4): 60–75. https://doi.org/10.4018/IJEHMC.20210701.oa4

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Website: https://www.r-project.org/ [accessed 3 April 2024].

Remita, M. A., A. Halioui, A. A. Malick Diouara, B. Daigle, G. Kiani, and A. B. Diallo. 2017. A machine learning approach for viral genome classification. *BMC Bioinformatics* 18(1): 208. https://doi.org/10.1186/s12859-017-1602-3

Sant'Anna, I. C., R. S. Tomaz, G. N. Silva, M. Nascimento, L. L. Bhering, and C. D. Cruz. 2015. Superiority of artificial neural networks for a genetic classification procedure. *Genetics and Molecular Research* 14(3): 9898–9906. https://doi.org/10.4238/2015.August.19.24

Saville, A., K. Graham, N. Grünwald, K. Myers, W. E. Fry, and J. Ristaino. 2015. Fungicide sensitivity of U.S. genotypes of *Phytophthora infestans* to six oomycete-targeted compounds. *Plant Disease* 99(5): 659–666. https://doi.org/10.1094/PDIS-05-14-0452-RE

Saville, A., F. La Spada, R. Faedda, Q. Migheli, B. Scanu, P. Ermacora, G. Gilardi, et al. 2021. Population structure of *Phytophthora infestans* collected on potato and tomato in Italy. *Plant Pathology* 70(9): 2165–2178. https://doi.org/10.1111/PPA.13444

Schiavo, G., F. Bertolini, G. Galimberti, S. Bovo, S. Dall'Olio, L. Nanni Costa, M. Gallo, and L. Fontanesi. 2020. A machine learning approach for the identification of population-informative markers from high-throughput genotyping data: Application to several pig breeds. *Animal* 14(2): 223–232. https://doi.org/10.1017/S1751731119002167

Shalev-Shwartz, S., O. Shamir, and E. Tromer. 2012. Using more data to speed-up training time. *In* N. D. Lawrence and M. Girolami [eds.], Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics, 1019–1027. Proceedings of Machine Learning Research. https://proceedings.mlr.press/v22/shalev-shwartz12.html

Sharma, A., A. Jain, P. Gupta, and V. Chowdary. 2021. Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access* 9: 4843–4873. https://doi.org/10.1109/ACCESS.2020.3048415

Tabima, J. F., S. E. Everhart, M. M. Larsen, A. J. Weisberg, Z. N. Kamvar, M. A. Tancos, C. D. Smart, et al. 2016. Microbe-ID: An open source toolbox for microbial genotyping and species identification. *PeerJ* 2016(8): e2279. https://doi.org/10.7717/peerj.2279

te Molder, D., W. Poncheewin, P. J. Schaap, and J. J. Koehorst. 2021. Machine learning approaches to predict the plant-associated phenotype of *Xanthomonas* strains. *BMC Genomics* 22(1): 848. https://doi.org/10.1186/s12864-021-08093-0

Torkzaban, B., A. H. Kayvanjoo, A. Ardalan, S. Mousavi, R. Mariotti, L. Baldoni, E. Ebrahimie, et al. 2015. Machine learning based classification of microsatellite variation: An effective approach for phylogeographic characterization of olive populations. *PLoS ONE* 10(11): e0143465. https://doi.org/10.1371/JOURNAL.PONE.0143465

Valdez-Valenzuela, E., A. Kuri-Morales, and H. Gomez-Adorno. 2021. Measuring the effect of categorical encoders in machine learning tasks using synthetic data. *Lecture Notes in Computer Science* 13067: 92–107. https://doi.org/10.1007/978-3-030-89817-5_7

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with S*, 4th ed. Springer, New York, New York, USA. https://doi.org/10.1007/978-0-387-21706-2

Wang, J., S. P. Fernández-Pavía, M. M. Larsen, E. Garay-Serrano, R. Gregorio-Cipriano, G. Rodríguez-Alvarado, N. J., Grünwald, and E. M. Goss. 2017. High levels of diversity and population structure in the potato late blight pathogen at the Mexico centre of origin. *Molecular Ecology* 26(4): 1091–1107. https://doi.org/10.1111/mec.14000

Warrens, M. J. 2011. Cohen's kappa is a weighted average. *Statistical Methodology* 8(6): 473–484. https://doi.org/10.1016/j.stamet.2011.06.002

Whisson, S. C., P. C. Boevink, S. Wang, and P. R. Birch. 2016. The cell biology of late blight disease. *Current Opinion in Microbiology* 34: 127–135. https://doi.org/10.1016/j.mib.2016.09.002

Zell, A., N. Mache, R. Hübner, G. Mamier, M. Vogt, M. Schmalzl, and K.-U. Herrmann. 2011. SNNS (Stuttgart Neural Network Simulator). *In* J. Skrzyppek [ed.], *Neural Network Simulation Environments*, 254, 165–186. Springer, New York, New York, USA. https://doi.org/10.1007/978-1-4615-2736-7_9

Zhao, J., G. Bodner, and B. Rewald. 2016. Phenotyping: Using machine learning for improved pairwise genotype classification based on root traits. *Frontiers in Plant Science* 7: 1864. https://doi.org/10.3389/fpls.2016.01864

How to cite this article: Patarroyo, C., S. Dupas, and S. Restrepo. 2024. A machine learning algorithm for the automatic classification of *Phytophthora infestans* genotypes into clonal lineages. *Applications in Plant Sciences* 12(5): e11603. https://doi.org/10.1002/aps3.11603