REVIEW

Application of machine learning in understanding plant virus pathogenesis: trends and perspectives on emergence, diagnosis, host-virus interplay and management

Dibyendu Ghosh[1†] , Srija Chakraborty[2†] , Hariprasad Kodamana[2,3] and Supriya Chakraborty[1*]

Abstract

Background: Inclusion of high throughput technologies in the field of biology has generated massive amounts of data in the recent years. Now, transforming these huge volumes of data into knowledge is the primary challenge in computational biology. The traditional methods of data analysis have failed to carry out the task. Hence, researchers are turning to machine learning based approaches for the analysis of high-dimensional big data. In machine learning, once a model is trained with a training dataset, it can be applied on a testing dataset which is independent. In current times, deep learning algorithms further promote the application of machine learning in several field of biology including plant virology.

Main body: Plant viruses have emerged as one of the principal global threats to food security due to their devastating impact on crops and vegetables. The emergence of new viral strains and species help viruses to evade the concurrent preventive methods. According to a survey conducted in 2014, plant viruses are anticipated to cause a global yield loss of more than thirty billion USD per year. In order to design effective, durable and broad-spectrum management protocols, it is very important to understand the mechanistic details of viral pathogenesis. The application of machine learning enables precise diagnosis of plant viral diseases at an early stage. Furthermore, the development of several machine learning-guided bioinformatics platforms has primed plant virologists to understand the host-virus interplay better. In addition, machine learning has tremendous potential in deciphering the pattern of plant virus evolution and emergence as well as in developing viable control options.

Conclusions: Considering a significant progress in the application of machine learning in understanding plant virology, this review highlights an introductory note on machine learning and comprehensively discusses the trends and prospects of machine learning in the diagnosis of viral diseases, understanding host-virus interplay and emergence of plant viruses.

Keywords: Machine learning, Deep learning, Plant virus, Pathogenesis, Host-virus interactions, Evolution and emergence, Control options

*Correspondence: schakraborty@mail.jnu.ac.in; supriyachakrasls@yahoo.com
†Dibyendu Ghosh and Srija Chakraborty have contributed equally to this article
[1] Molecular Virology Laboratory, School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India
Full list of author information is available at the end of the article

Background

Machine learning: an introduction for biologists

Over the years, extensive research has been carried out in various fields of biology to understand the science behind a plethora of complex biological phenomena. The study of problems such as traits in plants and plant

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 2 of 11

viral diseases lead to generation of massive data sets. The progress in technology has rendered data generation a simple task. Cost-effective technologies such as next generation sequencing (NGS) have made it easier to gather data regarding gene expression, chromosome conformation, genetic variation, traits and diseases of animals and plants, leading to generation of such massive data sets having multiple characteristics [1]. However, the resultant data explosion, especially in the field of omics, has made the handling of large datasets a major concern. The traditional statistical data analysis methodologies are not effective or efficient anymore in this context [2].

Furthermore, biological phenomena comprise various aspects, which lead to the generation of more than one data type. This necessitates an integrated analysis of the different types of data. But the noisiness of heterogeneous biological data makes this a difficult task [3]. Data dimensionality is another major impediment, for instance, omics data is generally highly resolved, hence highly dimensional. Moreover, the sample size in biological studies is limited in most cases. This may lead to issues including overfitting, multi-collinearity and data sparsity [4].

In order to overcome all these barriers, attempts are being made to incorporate machine learning (ML) and deep learning (DL) tools in the analysis of the datasets. ML tools identify patterns in the data using different statistical methods. Based on existing data, the ML paradigm can be used to derive models for classification, pattern recognition, and predictions. DL algorithms extract high-level features from huge datasets (such as collection of genomic sequences, or images), recognize the hidden patterns, and then utilize them to train models [5]. These trained models can be further applied to diverse types of unseen data from different sources for tasks such as prediction and classification. These techniques have the ability to tackle tough problems by detecting structure in seemingly random data, even when the amount of data is too complex and large for human comprehension [6]. Hence, ML especially DL, has the ability to perform analysis of enormous datasets in an extremely efficient, cost-effective, accurate and high-throughput manner [7].

In the context of ML, there are two primary frameworks for training the models: supervised and unsupervised learning. Both of these have potential for use in biology. Under supervised learning, the given collection features, or attributes of a system under investigation, are labeled [8]. Two recurring problems in the supervised learning framework are regression and classification. The classification process assigns objects into classes on the basis on the properties of features. In biology, one example of such training (involving mapping of object-to-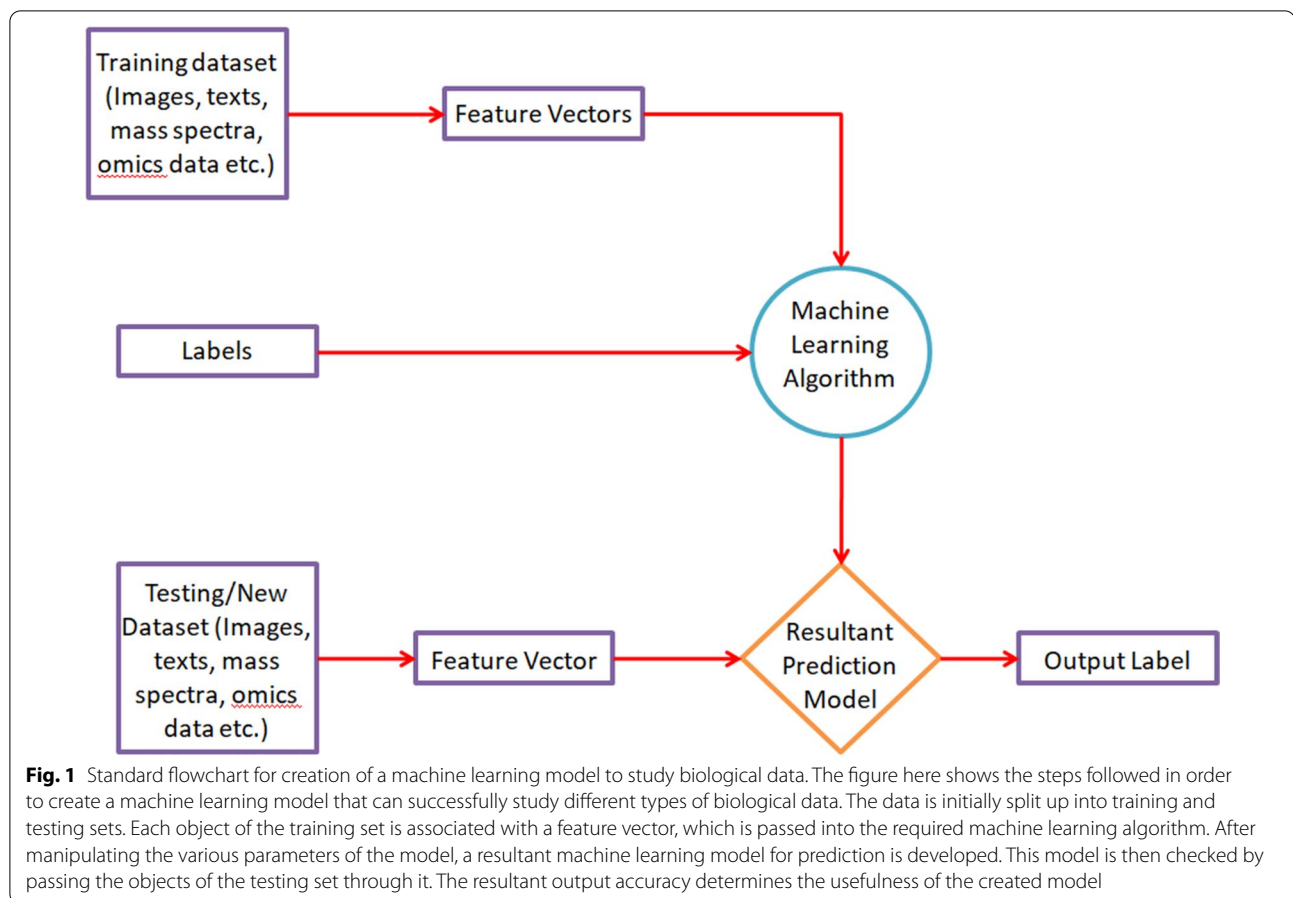class) is mapping of gene expression profiles to their respective diseases. The algorithm returns an assigned class of the object with certain "confidence measure" indicating the correctness of classification. Some of the widely used supervised models are linear/nonlinear regression, support vector machines (SVM), Gaussian processes, and neural nets [9].

In unsupervised learning, the objects involved in the study are not under any predefined labels [10]. The entire goal of these models is to recognize similarities in the various objects by exploring the data. These similarities define clusters in the data (groups of data objects). So, the basic concept of unsupervised learning is to discover natural patterns in data and categorize it into groups. Overall, in supervised learning, the data is pre-labeled and the algorithm learns how to utilize the labels to associate the objects to the classes. On the other hand, in unsupervised learning, the data is unlabeled, and the algorithm also learns to create labels by clustering the objects. Principal component analysis is an important example of unsupervised learning technique which includes k-means clustering, Gaussian mixture models, density-based spatial clustering of applications with noise (DBSCAN), and hierarchical clustering [9].

In certain exceptional cases, a method known as semi-supervised learning has proven to be quite useful. One example of such a scenario could be the classification of protein sequences. Only a few samples of protein sequences are labeled (belonging to a known class) but numerous sequences belong to unknown classes. The semi-supervised algorithm combines a small amount of labeled data with a large amount of unlabeled data during training [11].

The basic steps for creating a machine learning model for the study of biological data are shown in Fig. 1. Following the collection of data (labeled or unlabeled), it is divided into two sets for training and testing. The data samples need to undergo preprocessing and augmentation before the splitting in case they are corrupted with noise and outliers. Next the model is trained using the training dataset. The model can either be created from scratch, or a pre-trained model can be adjusted according to the collected dataset. Once the trained model is ready, the testing data is fed into it to determine the accuracy with which the objects are classified into different labels [12].

Deep learning is broad category of machine learning wherein large multiple layered neural network models are employed for representation learning. Deep learning can be performed in supervised, unsupervised, or semi-supervised fashion. When working with neural networks, we essentially attempt to create the inferences analogous to the human brain by building an artificial neural network (ANN) [13]. An ANN resembles a biological neural

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 3 of 11

Fig. 1 Standard flowchart for creation of a machine learning model to study biological data. The figure here shows the steps followed in order to create a machine learning model that can successfully study different types of biological data. The data is initially split up into training and testing sets. Each object of the training set is associated with a feature vector, which is passed into the required machine learning algorithm. After manipulating the various parameters of the model, a resultant machine learning model for prediction is developed. This model is then checked by passing the objects of the testing set through it. The resultant output accuracy determines the usefulness of the created model

network. The artificial neurons used here are basically mathematical models that carry out three main functions: activation, addition and multiplication. The goal is to build layers of neurons, each of which produces a suitable response to any input provided to it. The neurons of each layer multiply their inputs with the corresponding weights. Then it is passed through the activation function and finally transferred to the next layer of neurons. Once the input layer is fired up, the decision moves along to the subsequent layers of the neurons (hidden layers), firing up the respective neurons until the final output layer is reached [14]. A schematic representation of a neural network is presented in Fig. 2, where the various components of the network such as input, hidden and output layers are explained.

In neural networks, the direction of information flow is determined by the internodal connections. On this basis, there are two classifications of neural networks: (i) unidirectional flow which can further be divided into cascade forward and feed-forward; and (ii) bidirectional flow, also known as recurrent flow [15]. In feed-forward networks, the flow of information between the layers takes place in one direction. Cascade forward is similar except that the

input to the next layer is weighted. In recurrent networks, flow of information takes place in both directions. All the nodes are interconnected among each other, including self-connection. These networks are extremely complex, bulky, difficult to operate and take up a large amount of computational space. In addition to this, some neural networks architectures, such as self-organizing networks, convolutional neural networks (CNN), variational auto encoders (VAE) and generative adversarial networks (GAN) [16, 17] have recently attracted great attenttion in the DL community.

Various parameters are used to evaluate the classification performance of the developed model. Some of the important parameters include accuracy, sensitivity/recall rate, specificity rate, precision/positive predictive value, negative predictive value and $F1_{score}$ [18]. The performance of models is evaluated by calculating various ratios involving true positives, false positives, true negatives and false negatives. All of these can also be combined into a single confusion matrix to assess the model's performance [19]. Furthermore, the phenomena of overfitting and underfitting are widely faced while employing ML models [20]. Overfitting occurs when the model
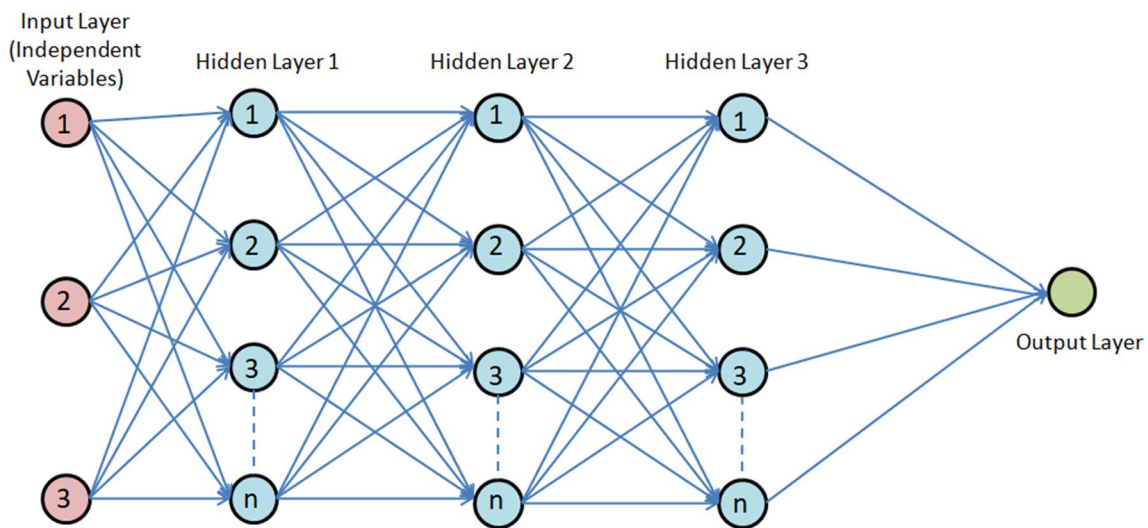
Ghosh *et al. Virology Journal*   (2022) 19:42

Page 4 of 11

Fig. 2 A schematic representation of a standard artificial neural network. The network is divided into three major components: the input layer, multiple hidden layers and the output layer. In this figure, it is assumed that the input layer has 3 independent variables, each of which is parsed through a set of weights and activation functions in the hidden layers and finally output layers to yield the model output. The activation functions are nonlinear mathematical function such as Tanh, Sigmoid, ReLU, etc. to induce nonlinearity to the model. Depending on the network structure, there may be 'n' neurons (also called hidden layer units) in each hidden layer and there may be multiple hidden layers. Any ANN with more than one hidden layer is technically is deep ANN. Once an input is fed into the network, one after another, each hidden layers gets operated among each other till finally the output layer is reached and activated, producing the final result. Weights in each layer is trained by means of the backpropagation algorithm

is fitted with respect to the noise in the data rather than the underlying signal. The validation data error increases while the training data error decreases [21]. On the other hand, underfitting is the reverse scenario. In this case, the model is not capable of recognizing data variability [22]. Several techniques including penalty methods, training by early stopping, batch normalization, and dropouts are being developed to avoid such imperfect fittings [23].

Main text

Application of machine learning in understanding plant viruses and viral infections

Diagnosis and detection of plant viral diseases

Plant viruses pose major economical constraints in cultivated crop plants across the world. Early detection of plant viral infection is crucial for successful disease management. An empirical evaluation through visible survey is traditionally followed by farmers to identify the symptoms of virus infected plants. The visual assessment bias dictates the inefficiency and inaccuracy of this method. On the other hand, laboratory-based detection techniques are primarily reliant on polymerase chain reaction (PCR) and serological-based method such as enzyme linked immune sorbent assay (ELISA). Despite their improved accuracy, the requirement for professional experts and their time-consuming and invasive nature pinpoints the shortcomings of these diagnosis assays [24,

25]. A pathogen attack significantly alters the biochemical and biophysical state of the plant leading to an alteration of tissue structure, water level and transpiration rate, ultrastructure of chloroplast and pigment content [26, 27]. At the very beginning of the twenty-first century, a few studies used remote sensors to capture and detect altered leaf reflectance and thermography profiling of diseased plants, which empowered the scientific community with an edge in phenotyping of stressed plants [28]. However, this technique was unable to determine if the stress was biotic or abiotic, and if biotic, what kind of pathogen was involved. Hyper-spectral imaging (HSI) and ML assisted data analysis are now revolutionizing the concept of stress phenotyping of diseased plants by enabling the diagnosis of specific plant diseases and even the severity of the disease. In the case of HSI, a light spectrum with a larger range of wavelengths is being used to capture plant images, which enables us to go beyond the limited range of human vision (400-700 nm) in monitoring minor alterations in the growth and development of plants [29]. For ML assisted detection of plant viral diseases, first a ML model has to be trained with a training dataset (images of diseased plants captured through unmanned aerial vehicle, grounded robots or even smartphones) [30–38]. There are increasing numbers of free online databases which provide images of specific plant diseases as training datasets. 'Plantvillage' is one

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 5 of 11

such initiative [39]. Once a certain ML model has been trained accurately and precisely, a testing dataset (eg: hyperspectral images of specific plants under diagnosis) can be assessed [32, 34]. HSI generates high dimensional data with redundant information and hence, an efficient pre-processing of the data is crucial for the precise functioning of the model. An effective specific range of wavelength can be determined to reduce the dimensionality of HSI data [29]. The next step is feature extraction which minimizes the number of the features present in the raw dataset. The feature extraction method is vital for assuring a simple classifier with a limited variety of features, since multifeatured classification always hinders the smooth performance of the concerned model [40, 41]. ML researchers have devised a variety of feature extraction techniques based on the nature of the data and the model. However, the process is time-consuming and the success of the operation greatly relies on the expertise of the professional. Here comes the benefits of using DL techniques for feature extraction as DL empowers automatic extraction of features rather than handcrafted method used in traditional ML algorithms, for instance the application of convolutional neural networks. DL has substantially improved the reliability of plant stress phenotyping by enabling the accommodation of a large sample size for training and testing [40]. A major constraint of this method is the vast variation of environmental conditions between the field and the lab. While consistent temperature, humidity, and light intensity are maintained in the lab, all of these variables are constantly changing in the field, influencing the captured images [42]. Hence, it is recommended to use field images to train a model since it has been demonstrated that a classifier trained on field images can also classify lab-based images with precision [43]. The lack of availability of a huge collection of field-based images of a specific plant disease is another key challenge for an accurate and reliable diagnosis.

Transfer learning is a recent advancement in the field of ML which enables the data scientists to adopt a previously well-trained model for solving a similar kind of problems [40]. For example-a model trained for chilli-leaf curl disease detection may be used for detecting leaf curl symptoms caused by viruses in tomato (Fig. 3). There are several approaches to adopt a pre-trained model; one can select and finetune the architecture and/or parameters of a model depending upon the types of datasets. Table 1 summarizes the development of ML assisted diagnosis of plant viral diseases over last few years.

Understanding the diversity and emergence of plant viruses

The recent trend of studying plant virome through metagenomics has unveiled the diversity of plant viruses. Huge numbers of phylogenetically related and unrelated virus species have been found in diseased samples [44, 45]. Explosion in virome data generated through NGS necessitates the urgent structural orientation and analysis of sequence data in order to understand the actual portrait of the viral diversity. Although a significant progress has been followed up in the case of animal viruses, limited efforts have yet been recorded in the field of plant virology [46, 47]. V-pipe has provided a bioinformatics pipeline for analyzing genomic diversity of human immunodeficiency virus (HIV) from sequencing data [48].

As RNA viruses use error-prone polymerases during their replication, the chances of mutations in their genome sequences remain quite high. Mutation in the viral genome finally leads to the emergence of new virulent viral strains [49]. A neural network-based model can predict probable point mutations in the RNA sequence. It has been successfully explored in the case of newcastle virus [50], and its optimized form may be very useful for the prediction of mutations in plant viral genome (Fig. 3). Besides RNA viruses, DNA viruses also possess significant genetic variations. Events like recombination and genome reassortment play crucial role in mediating the emergence of new viral forms [51, 52]. The identification of novel virus and satellite molecules through metagenomics approach emphasizes the importance of precise taxonomic classification followed by demarcation of these new species. An excellent effort by Silva and collaborators have developed Fangorn Forest, a ML based method, for classification of geminiviruses. Among the three tested algorithms, random forest (RF) has proven to be best in classification of genes and genera of this largest plant virus family [53]. Recently, a CNN guided sequencing platform has successfully completed human genome sequencing within couple of hours and efficiently identified the disease-causing variations in the genome. This ML based fastest sequencing approach may open up new windows in studying diversity and evolution of plant viruses [54].

Understanding host-virus interplay

Being obligate parasites, viruses rely on cellular machineries of plants for every aspect of pathogenesis including replication, gene expression and movement [55]. Plants elicit a robust antiviral immune response to restrict viral invasion [56]. Viruses encode effector proteins which disarm plant defense signaling. This tug of war continues which fuels the co-evolution of both virus and host [57]. Hence, understanding the interplay between plant and viruses is crucial for an in-depth dissection of viral pathogenesis.

Although plants have evolved a variety of tools and tactics to prevent virus multiplication, the resistance (R) protein-mediated immune response and gene silencing
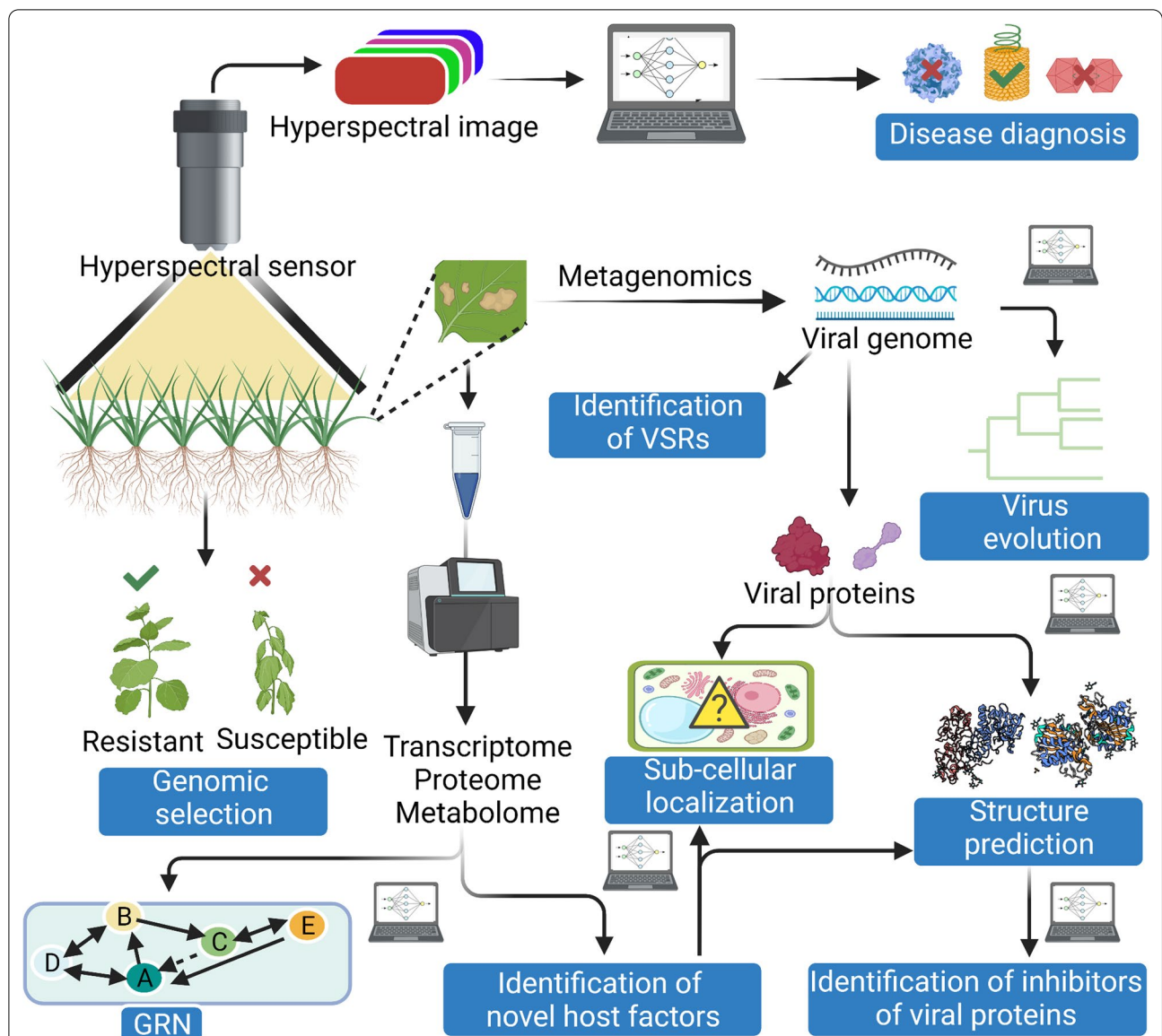
Fig. 3 Application of ML in understanding plant virus pathogenesis. ML enables early diagnosis of plant viral diseases at field level through analyzing hyperspectral images. Metagenomics study of diseased plant samples helps identification of related and unrelated viral genomes. ML can assist in the classification of these viral sequences which primes our understanding of virus evolution. Furthermore, ML-assisted bioinformatics tools have been developed to identify viral suppressors of RNA silencing (VSRs). ML can also guide us to predict the sub-cellular localization and even the structure of the viral proteins. Prediction of accurate structures of virus encoded proteins may help to identify inhibitors of these effector proteins. To understand the host response, several groups have performed transcriptome, proteome and metabolome of virus infected plants. ML can prime the accurate and fast analysis of these high throughput data to identify gene regulatory networks (GRN) and novel host factors involved in host-virus interplay. Characterization of these host factors in terms of sub-cellular localization and structure prediction will boost understanding of plant virus pathogenesis. ML may also assist plant virologists in genomic selection to identify elite virus resistant cultivars. This figure was created using BioRender (https://biorender.com/)

are the most well-known features of their antiviral defense [56]. A majority of canonical R-proteins contain nucleotide binding site leucine-rich repeats (NBS-LRR), which mediate direct or indirect recognition of virus-encoded effector proteins, resulting in the activation of effector triggered immunity (ETI). Very few *R*-genes

imparting immune response against viruses have been identified and characterised till date, which limits our knowledge regarding the detailed mechanism of dominant resistance in plant virus interaction [58]. Support vector machine-assisted development of a high throughput bioinformatics tool, NBSPred, precisely identifies

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 7 of 11

Table 1 The application of ML-assisted diagnosis of plant viral diseases

Plant	Viruses/viral diseases	Algorithms/methodologies used	Accuracy	References
Cassava	Cassava mosaic disease	Convolutional neural networks (CNN)	96%	[31]
	Cassava brown streak disease		98%	
Cucumber	(i) Melon yellow spot virus; (ii) Zucchini yellow mosaic virus	CNN	94.9%	[30]
Mungbean	Yellow mosaic disease	CNN	91.234% for VirLeafNet-1	[37]
			96.429% for VirLeafNet-2	
			97.403% for VirLeafNet-3	
Potato	Potato virus Y	Support vector machine (SVM) classifier	89.8%	[36]
Sweet pepper	Tomato spotted wilt virus (TSWV)	Outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN)	96.25% (before the onset of visible symptoms)	[32]
Tobacco	Tobacco mosaic virus (TMV)	Successive projections algorithm (SPA) with extreme learning machine (ELM) classifier	98.33%	[34]
Tobacco	TMV	SVM	93.5% on the training set	[41]
			92.7% on the independent set	
Tobacco	TSWV	Model by boosted regression tree (BRT) algorithm and Wavelength selection by SPA	85.2%	[35]
Tobacco	Tomato leaf curl New Delhi virus and Tomato leaf curl Gujarat virus	CNN [Visual Geometry Group 16]	97.21%	[38]
Tomato	Groundnut bud necrosis virus (GBNV)	SVM	97.8%	[33]

NBS-LRR containing R proteins from genome, transcriptome and proteome data [59]. Receptor-like kinases (RLK) are crucial players in the immune perception of phytopathogens, many of them acting as pattern recognition receptors (PRRs) which lead to induction of pattern triggered immunity (PTI) [60]. However, several plant viruses target RLKs to promote viral pathogenesis [61]. Brustolini et al. have recently developed a machine learning assisted technique for detection of RLKs from proteome data. Identification and annotation of novel RLKs may advance our current understanding of plant-virus interactions [62]. Furthermore, to identify host factors differentially regulated in host-virus interplay, several groups have performed transcriptome and proteome analysis in both resistant and susceptible plant varieties. These studies have revealed that a significant proportion of differentially expressed transcripts are of unknown nature suggesting the existence of novel gene regulatory networks (GRNs) modulating the host-virus interaction [63–66]. ML helps biologists to predict GRNs from high-throughput transcriptome data [67] which may lead to identification of several regulatory nodes of plant immune signalling (Fig. 3).

On the other hand, viruses encode few but multitasking effector proteins which facilitate the viral pathogenesis. Examining the sub cellular localization of these effector proteins is important to understand their mechanism of action. Furthermore, viruses also redirect the subcellular localization of several host proteins to disrupt their assigned functions [68]. ML assisted development of online tools such as LOCALIZER and MU-LOC enable precise as well as accurate analysis of subcellular localization of effector proteins and host factors by simply using amino acid sequences of proteins as input (Fig. 3) [69, 70]. Application of ML in the successful prediction of fungal effector proteins has added an extra edge in phytopathology research [71]. In the case of viruses, some viral effector proteins have been evolved to block antiviral gene silencing, known as viral suppressors of RNA silencing (VSRs). VSRs expand the negative impact of viral diseases by promoting synergistic associations among different plant viruses [72]. Jagga et al. have developed a bioinformatics platform, pVsupPred, for the prediction of VSRs encoded by plant-infecting viruses. They have used four classifier models including LibSVM, J48, Naïve Bayes and RF, and among all of them, RF algorithm has emerged as the best with an overall accuracy of 86.11% [73]. Later on, in another study, sequential minimal optimization (SMO) algorithm had been improved to achieve an overall accuracy of 95.3% for the successful identification of plant virus encoded VSRs (Fig. 3) [74].

Another significant facet of plant-virus interaction is the virus induced alteration of microRNA (miRNA) homeostasis which impacts the transcriptome profile of the infected cells. Hence, it is important to identify the accurate targets of specific miRNAs regulating plant immunity and viral pathogenesis [75]. The advent of ML in advancing the scope of bioinformatics has significantly eased this difficult job. Supervised ML approaches including graphical models, kernel machines and

Table 2 Brief description of ML-based bioinformatics platforms used in studying plant virus interactions

Name	Application	Input and output	Salient features	References
V-PIPE	Assess genetic diversity of viral population and ensure identification of true viral variants from high throughput data	**Input:** raw sequencing data (FASTQ format) **Output:** viral diversity in terms of single nucleotide variants, local and global viral haplotypes	A hidden Markov model-based read aligner, ngshmmalign, is developed	[48]
NBSPred	Identify potential NBS-LRR and NBS-LRR like proteins	**Input:** Genome, transcripts and protein sequences **Output:** Identification of NBS-LRR and NBS-LRR like proteins	Gene prediction tool, Augustus2.7, is used to convert genomic sequences to protein sequences. TransDecoder is used to convert transcripts sequences to protein sequences (i) Frequency of aminoacids, dipeptides, tripeptides and multiplet; (ii) charge (iii) hydrophobicity are considered for the calculation of sequence compositional property	[59]
LOCALIZER	Predict the sub cellular localization of plant proteins and effector proteins encoded by plant-infecting fungus and oomycete	**Input:** sequence of plant proteins and eukaryotic effector proteins **Output:** (i) probability of localization of a protein in nucleus, chloroplast or mitochondria (ii) Identification of transit peptides (for chloroplast and mitochondria) and nuclear localization signal (NLS)	Trained by support vector machine model Maximum range: 2000 sequences	[69]
MU-LOC	Predict the mitochondrial localization of plant proteins	**Input:** protein sequence (FASTA format) **Output:** sub cellular localization	The predictor has been trained using support vector machine and deep neural network	[70]
pVsupPred	Predict RNA silencing suppressor activity of viral proteins (VSR)	**Input:** sequence of viral proteins **Output:** (i) prediction score, (ii) Whether positive VSR or negative VSR	Random forest model guided tool Prediction on the basis of presence of (i) GW/WG motif and (ii) dsRNA binding domain in the viral protein	[73]
Alphafold	Predict the structure of a protein	**Input:** amino acid sequence of a protein **Output:** 3D structure of the protein	Neural-network based model Median accuracy: (i) 6.6 Å for Alphafold, (ii) 1.5 Å for Alphafold2	[78]
Virfinder	Identify sequences of viruses from metagenomic data	**Input:** assembled metagenomic data **Output:** true viral contigs	*k*-mer based prediction tool has been made using a trained logistic regression model	[81]

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 9 of 11

evolutionary algorithms are being widely used to identify the specific miRNA targets in eukaryotes [76]. Further, a new category of DL models known as graph neural nets (GNN) is emerging as a promising tool in bioinformatics. The biological networks, based on small RNAs–disease associations, can be constructed as graphs with nodes and edges. GNN can operate on the graphical data and has more representative features, which can be efficiently used for inferences [77].

Finally, the best possible way to understand the functional aspect of a protein is to visualize its accurate structure. A very small proportion of plant proteins involved in immune signalling have been structurally characterised yet. In addition, structures of plant viral proteins are also largely unresolved. Labour intensive methods of protein crystallization is the major bottleneck here. However, Jumper and collaborators have revolutionised the idea of protein structure prediction by launching Alphafold2, a neural network-assisted structural bioinformatics platform, which can successfully solve a protein structure with almost equivalent experimental accuracy even if there is no similar protein structure available [78]. ML-guided docking studies efficiently screen chemical inhibitors of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) encoded spike (S) protein [79]. Similarly, structure prediction of plant viral proteins and the prediction of their chemical inhibitors followed by successful delivery will be a novel and effective virus management strategy (Fig. 3).

Conclusions and future perspectives

Plant-infecting viruses not only compromise the yield of the infected plants but also significantly affect the nutritional content of crops. Heavy crop losses due to plant viral outbreaks is a vital concern for global food security and hence necessitates the urgent implementation of smart management measures. Studies aiming to understand the evolutionary biology of plant viruses and the molecular biology of plant-virus interactions have generated large-sized datasets in recent years. Here comes the prospective role of ML. Although the last decade has witnessed a sharp increase of application of ML in solving complex biological problems [80], its usage in the field of plant virology is still at a very naïve state. Several reports highlighted the role of ML in the precise diagnosis of plant viral diseases [30–38]. Plant virologists can foresee the tremendous scope of ML in addressing virus evolution, emergence, plant-virus interplay and above all management strategies (Table 2). Moreover, several specific issues need to be explored.

Firstly, a vast amount of OMICS data (including transcriptome, proteome and metabolome) of virus infected plants are available. The application of ML

may enable the integration of these OMICS data which will definitely uplift our knowledge of host response, especially the impact of novel potential host factors in viral infection. Secondly, the coordinated application of ML and HSI pave a new path in the detection of viral diseases. Now, designing improved DL-based algorithms for easily accessible mobile app-mediated detection of plant viral diseases is the need of the hour. Thirdly, ML algorithms would be very helpful in deciphering the patterns and parameters of plant viral evolution. ML may enable the accurate prediction of recombination, rate of nucleotide substitutions, mutations and phylogenetic relatedness among viruses. The use of ML to predict the advent of aggressive recombinant plant virus strains and the likelihood of associated epidemics will be extremely beneficial. Fourthly, a significant improvement can be achieved in metagenomics data analysis through ML approaches. Along with viral reads, metagenomics data also contains a substantial proportion of host contig contaminations which often hinders the identification of small viral reads. VirFinder is one such *k*-mer based platform which enables the identification of prokaryotic virus sequences from mixed metagenomic data [81]. A similar approach can be employed for plant virome study. Fifthly, ML is now widely utilized in genomic selection for rapid and better prediction of superior genotypes for breeding purposes [82, 83]. A certain progress of ML assisted genomic prediction will definitely help breeders in developing elite virus tolerant/ resistant varieties. Finally, a collaborative effort by both plant virologists and big data analysts is of prime importance for the fruitful application of ML in the understanding of plant virus pathogenesis followed by the development of antiviral strategies.

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 10 of 11

Availability of data and materials
Not applicable.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
All authors have agreed for publication in the journal.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Molecular Virology Laboratory, School of Life Sciences, Jawaharlal Nehru University, New Delhi 110067, India. [2]Department of Chemical Engineering, Indian Institute of Technology Delhi, New Delhi 110016, India. [3]School of Artificial Intelligence, Indian Institute of Technology Delhi, New Delhi 110016, India.

References
1. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferrero E, Agapow P-M, Zietz M, Hoffman MM, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15:20170387.
2. Xu C, Jackson SA. Machine learning and complex biological data. Genome Biol. 2019;20:76.
3. Zitnik M, Nguyen F, Wang B, Leskovec J, Goldenberg A, Hoffman MM. Machine learning for integrating data in biology and medicine: principles, practice, and opportunities. Information Fusion. 2019;50:71–91.
4. Altman N, Krzywinski M. The curse(s) of dimensionality. Nat Methods. 2018;15:399–400.
5. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018;15:233–4.
6. Webb S. Deep learning for biology. Nature. 2018;554:555–8.
7. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521:436–44.
8. Singh A, Ganapathysubramanian B, Singh AK, Sarkar S. Machine learning for high-throughput stress phenotyping in plants. Trends Plant Sci. 2016;21:110–24.
9. Tarca AL, Carey VJ, Chen X, Romero R, Drăghici S. Machine learning and Its applications to biology. PLoS Comput Biol. 2007;3:e116.
10. Prasad V, Gupta SD. Applications and potentials of artificial neural networks in plant tissue culture. In: Plant tissue culture engineering. Springer; 2008. p. 47–67
11. Weston J, Leslie C, Ie E, Zhou D, Elisseeff A, Noble WS. Semi-supervised protein classification using cluster kernels. Bioinformatics. 2005;21:3241–7.
12. Tang B, Pan Z, Yin K, Khateeb A. Recent advances of deep learning in bioinformatics and computational biology. Front Genet. 2019;10:214.
13. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. Nature. 2015;518:529–33.
14. Krenker A, Bešter J, Kos A. Introduction to the artificial neural networks. Artificial neural networks: methodological advances and biomedical applications. InTech 2011. pp. 1–18.
15. Osama K, Mishra BN, Somvanshi P: Machine learning techniques in plant biology. In PlantOmics: The omics of plant science. Springer; 2015. Pp. 731–54.
16. Yao X. Evolving artificial neural networks. Proc IEEE. 1999;87:1423–47.
17. Yang ZR. A novel radial basis function neural network for discriminant analysis. IEEE Trans Neural Netw. 2006;17:604–12.
18. Taner A, Öztekin YB, Duran H. Performance analysis of deep learning CNN models for variety classification in hazelnut. Sustainability. 2021;13:6527.
19. Hassan SM, Maji AK, Jasiński M, Leonowicz Z, Jasińska E. Identification of plant-leaf diseases using CNN and transfer-learning approach. Electronics. 2021;10:1388.
20. Smith LN. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay. arXiv preprint arXiv:1803.09820. 2018.
21. Sarle WS. Stopped training and other remedies for overfitting. Comput Sci Stat. 1996;66:352–60.
22. Van der Aalst WM, Rubin V, Verbeek H, van Dongen BF, Kindler E, Günther CW. Process mining: a two-step approach to balance between underfitting and overfitting. Softw Syst Model. 2010;9:87–111.
23. Murakoshi K. Avoiding overfitting in multilayer perceptrons with feeling-of-knowing using self-organizing maps. BioSystems. 2005;80:37–40.
24. Rubio L, Galipienso L, Ferriol I. Detection of plant viruses and disease management: relevance of genetic diversity and evolution. Front Plant Sci. 2020;11:1092.
25. Varma A, Singh MK. Chapter 6—Diagnosis of plant virus diseases. In: Awasthi LP, editor. Applied plant virology. Academic Press; 2020. p. 79–92.
26. Bhattacharyya D, Gnanasekaran P, Kumar RK, Kushwaha NK, Sharma VK, Yusuf MA, Chakraborty S. A geminivirus betasatellite damages the structural and functional integrity of chloroplasts leading to symptom formation and inhibition of photosynthesis. J Exp Bot. 2015;66:5881–95.
27. Pallas V, García JA. How do plant viruses induce disease? Interactions and interference with host components. J Gen Virol. 2011;92:2691–705.
28. Landgrebe DA. Signal theory methods in multispectral remote sensing. Wiley; 2003.
29. Lowe A, Harrison N, French AP. Hyperspectral image analysis techniques for the detection and classification of the early onset of plant disease and stress. Plant Methods. 2017;13:80.
30. Kawasaki Y, Uga H, Kagiwada S, Iyatomi H: Basic study of automated diagnosis of viral plant diseases using convolutional neural networks. In: International symposium on visual computing. Springer; 2015. pp. 638–45.
31. Ramcharan A, Baranowski K, McCloskey P, Ahmed B, Legg J, Hughes DP. Deep learning for image-based cassava disease detection. Front Plant Sci. 1852;2017:8.
32. Wang D, Vinson R, Holmes M, Seibel G, Bechar A, Nof S, Tao Y. Early detection of tomato spotted wilt virus by hyperspectral imaging and outlier removal auxiliary classifier generative adversarial nets (OR-AC-GAN). Sci Rep. 2019;9:1–14.
33. Kadam KU. Identification of groundnut bud necrosis virus on tomato fruits using machine learning based segmentation algorithm. Int J Fut Gener Commun Netw. 2020;13:259–64.
34. Zhu H, Chu B, Zhang C, Liu F, Jiang L, He Y. Hyperspectral imaging for presymptomatic detection of tobacco disease with successive projections algorithm and machine-learning classifiers. Sci Rep. 2017;7:1–12.
35. Gu Q, Sheng L, Zhang T, Lu Y, Zhang Z, Zheng K, Hu H, Zhou H. Early detection of tomato spotted wilt virus infection in tobacco using the hyperspectral imaging technique and machine learning algorithms. Comput Electron Agric. 2019;167:105066.
36. Griffel L, Delparte D, Edwards J. Using support vector machines classification to differentiate spectral signatures of potato plants infected with Potato Virus Y. Comput Electron Agric. 2018;153:318–24.
37. Joshi RC, Kaushik M, Dutta MK, Srivastava A, Choudhary N. VirLeafNet: Automatic analysis and viral disease diagnosis using deep-learning in Vigna mungo plant. Ecol Inf. 2021;61:101197.
38. Chakraborty S, Kodamana H, Chakraborty S. Deep learning aided automatic and reliable detection of tomato begomovirus infections in plants. J Plant Biochem Biotechnol. 2021;66:1–8.

Ghosh *et al. Virology Journal*   (2022) 19:42

Page 11 of 11

39. Gao Z, Luo Z, Zhang W, Lv Z, Xu YJA. Deep learning application in plant stress imaging: a review. AgriEngineering. 2020;2:430–46.

40. Singh AK, Ganapathysubramanian B, Sarkar S, Singh A. Deep Learning for plant stress phenotyping: trends and future perspectives. Trends Plant Sci. 2018;23:883–98.

41. Chen Y-M, Zu X-P, Li D. Identification of proteins of Tobacco mosaic virus by using a method of feature extraction. Front Genet. 2020;11:1186.

42. Mahlein A-K. Plant disease detection by imaging sensors—parallels and specific demands for precision agriculture and plant phenotyping. Plant Dis. 2016;100:241–51.

43. Ferentinos KP. Deep learning models for plant disease detection and diagnosis. Comput Electron Agric. 2018;145:311–8.

44. Roossinck MJ. Plant virus metagenomics: biodiversity and ecology. Annu Rev Genet. 2012;46:359–69.

45. Stobbe AH, Roossinck MJ. Plant virus metagenomics: what we know and why we need to know more. Front Plant Sci. 2014;5:150.

46. Li J, Zhang S, Li B, Hu Y, Kang X-P, Wu X-Y, Huang M-T, Li Y-C, Zhao Z-P, Qin C-F. Machine learning methods for predicting human-adaptive influenza A viruses based on viral nucleotide compositions. Mol Biol Evol. 2020;37:1224–36.

47. Randhawa GS, Soltysiak MP, El Roz H, de Souza CP, Hill KA, Kari L. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. PLoS ONE. 2020;15:e0232391.

48. Posada-Céspedes S, Seifert D, Topolsky I, Jablonski KP, Metzner KJ, Beerenwinkel N. V-pipe: a computational pipeline for assessing viral genetic diversity from high-throughput data. Bioinformatics. 2021;6:66.

49. Elena SF, Agudelo-Romero P, Carrasco P, Codoñer FM, Martín S, Torres-Barceló C, Sanjuán R. Experimental evolution of plant RNA viruses. Heredity. 2008;100:478–83.

50. Salama MA, Hassanien AE, Mostafa A. The prediction of virus mutation using neural networks and rough set techniques. EURASIP J Bioinf Syst Biol. 2016;2016:1–11.

51. Kumar RV, Singh AK, Singh AK, Yadav T, Basu S, Kushwaha N, Chattopadhyay B, Chakraborty S. Complexity of begomovirus and betasatellite populations associated with chilli leaf curl disease in India. J Gen Virol. 2015;96:3143–58.

52. Devendran R, Kumar M, Ghosh D, Yogindran S, Karim MJ, Chakraborty S. Capsicum-infecting begomoviruses as global pathogens: host-virus interplay, pathogenesis, and management. Trends Microbiol. 2021;6:66.

53. Silva JCF, Carvalho TFM, Fontes EPB, Cerqueira FR. Fangorn Forest (F2): a machine learning approach to classify genes and genera in the family Geminiviridae. BMC Bioinformatics. 2017;18:431.

54. Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, Spiteri E, Pesout T, Monlong J, Baid G, et al. Ultrarapid nanopore genome sequencing in a critical care setting. New Engl J Med. 2022;6:66.

55. Mandadi KK. Scholthof K-BG: Plant immune responses against viruses: How does a virus cause disease? Plant Cell. 2013;25:1489–505.

56. Calil IP, Fontes EPB. Plant immunity against viruses: antiviral immune receptors in focus. Ann Bot. 2017;119:711–23.

57. Wu X, Valli A, García JA, Zhou X, Cheng X. The tug-of-war between plants and viruses: great progress and many remaining questions. Viruses. 2019;66:11.

58. Kourelis J, van der Hoorn RAL. Defended to the Nines: 25 years of resistance gene cloning identifies nine mechanisms for R protein function. Plant Cell. 2018;30:285–99.

59. Kushwaha SK, Chauhan P, Hedlund K, Ahrén D. NBSPred: a support vector machine-based high-throughput pipeline for plant resistance protein NBSLRR prediction. Bioinformatics. 2015;32:1223–5.

60. Tang D, Wang G, Zhou JM. Receptor kinases in plant-pathogen interactions: more than pattern recognition. Plant Cell. 2017;29:618–37.

61. Macho AP, Lozano-Duran R. Molecular dialogues between viruses and receptor-like kinases in plants. Mol Plant Pathol. 2019;20:1191–5.

62. Brustolini OJ, Silva JC, Sakamoto T, Fontes EP. Bioinformatics analysis of the receptor-like kinase (RLK) superfamily. Methods Mol Biol. 2017;1578:123–32.

63. Liu D, Zhao Q, Cheng Y, Li D, Jiang C, Cheng L, Wang Y, Yang A. Transcriptome analysis of two cultivars of tobacco in response to Cucumber mosaic virus infection. Sci Rep. 2019;9:3124.

64. Liu Y, Liu Y, Spetz C, Li L, Wang X. Comparative transcriptome analysis in *Triticum aestivum* infecting wheat dwarf virus reveals the effects of viral infection on phytohormone and photosynthesis metabolism pathways. Phytopathol Res. 2020;2:3.

65. Rajamäki M-L, Sikorskaite-Gudziuniene S, Sarmah N, Varjosalo M, Valkonen JPT. Nuclear proteome of virus-infected and healthy potato leaves. BMC Plant Biol. 2020;20:355.

66. Sade D, Shriki O, Cuadros-Inostroza A, Tohge T, Semel Y, Haviv Y, Willmitzer L, Fernie AR, Czosnek H, Brotman Y. Comparative metabolomics and transcriptomics of plant response to Tomato yellow leaf curl virus infection in resistant and susceptible tomato cultivars. Metabolomics. 2015;11:81–97.

67. Mochida K, Koda S, Inoue K, Nishii R. Statistical and machine learning approaches to predict gene regulatory networks from transcriptome datasets. Front Plant Sci. 2018;9:1770.

68. Rodriguez-Peña R, Mounadi KE, Garcia-Ruiz H. Changes in subcellular localization of host proteins induced by plant viruses. Viruses. 2021;13:677.

69. Zhang N, Rao RSP, Salvato F, Havelund JF, Møller IM, Thelen JJ, Xu D. MU-LOC: a machine-learning method for predicting mitochondrially localized proteins in plants. Front Plant Sci. 2018;9:634.

70. Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci Rep. 2017;7:44598.

71. Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, Manners JM, Taylor JM. EffectorP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 2016;210:743–61.

72. Ghosh D. M M, Chakraborty S: Impact of viral silencing suppressors on plant viral synergism: a global agro-economic concern. Appl Microbiol Biotechnol. 2021;105:6301–13.

73. Jagga Z, Gupta D. Supervised learning classification models for prediction of plant virus encoded RNA silencing suppressors. PLoS ONE. 2014;9:e97446.

74. Nath A, Subbiah K. Probing an optimal class distribution for enhancing prediction and feature characterization of plant virus-encoded RNA-silencing suppressors. 3 Biotech. 2016;6:93.

75. Zhang B, Li W, Zhang J, Wang L, Wu J. Roles of small RNAs in virus-plant interactions. Viruses. 2019;11:827.

76. Zhang B-T, Nam J-W. Supervised learning methods for microRNA studies. In: Machine learning in bioinformatics. 2008. pp 339–365.

77. Zhang X-M, Liang L, Liu L, Tang M-J. Graph neural networks and their current applications in bioinformatics. Front Genet. 2021;12:690049.

78. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.

79. Batra R, Chan H, Kamath G, Ramprasad R, Cherukara MJ, Sankaranarayanan SK. Screening of therapeutic agents for COVID-19 using machine learning and ensemble docking studies. J Phys Chem Lett. 2020;11:7058–65.

80. Ma C, Zhang HH, Wang X. Machine learning for big data analytics in plants. Trends Plant Sci. 2014;19:798–808.

81. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.

82. Poland J, Rutkoski J. Advances and challenges in genomic selection for disease resistance. Annu Rev Phytopathol. 2016;54:79–98.

83. Crossa J, Pérez-Rodríguez P, Cuevas J, Montesinos-López O, Jarquín D, de los Campos G, Burgueño J, González-Camacho JM, Pérez-Elizalde S, Beyene Y, et al: Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 2017;22:961–75.

Publisher's Note