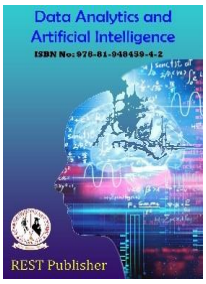
Data Analytics and Artificial Intelligence
Vol: 3(7), 2023
REST Publisher; ISBN: 978-81-948459-4-2
Website: http://restpublisher.com/book-series/daai/



PRED-PATHO: Prediction of Pathogenesis Related Protein in Plants using Machine Learning Algorithms

*A. Amuthavalli, N. Priya
Shrimathi Devkunvar Nanalal Bhatt Vaishnav College for Women
University of Madras, Chennai, Tamil Nadu, India.
*Corresponding Author Email: a.s.amuthavalli@gmail.com

Abstract: *Biotic stress is one of major elements that constantly affect plant growth and survival. In response to various biotic stress conditions like bacteria, virus or fungi attack, plants trigger their defence system, which includes the expression of a number of proteins, signalling molecules, phytochemicals. One of the proteins that is most frequently activated during biotic stress is the pathogenesis related (PR) protein. They are essential elements of the plant immune system, and are frequently used as biochemical indicators of signalling pathways that support defenceagainst fungi, bacteria and viruses attack. With advent of enhanced high throughput genomics and proteomics data, machine learning technologies are playing a crucial role in predicting potential biotic stress related genes and its interaction with various biological networks in protecting plants against various pathogens.This study utilized machine learning methodologies to forecast and analyze PR proteins across diverse plant species. The study examined PR in the chosen plant protein database and showed that machine learning models trained with sizable amounts of essentially favourable data sets could analyse PR much more effectively than the current methods. Employing a combination of machine learning techniques such as Random Forest (RF), Naive Bayes (NB), and K-Nearest Neighbors (KNN), the development of the proposed PRED-PATHO system resulted in classification accuracies of 94% for PR using RF, 82% using NB, and 71% using KNN.*

Keywords: *Machine learning, Biotic Stress, Random Forest, Naïve bayes, KNN, Classification.*

1. INTRODUCTION

Plants are continually threatened by a variety of biotic factors like pathogenic bacteria, fungus, and viruses, which affects plant growth and development [1]. These infections significantly reduce agricultural revenues each year and extremely jeopardise the future security of our food supply [2]. In order to live or maintain their fitness, plants employ a variety of defence mechanisms against these adversaries[3]. At event of biotic stress, plant produce a kind of defense related signalling molecule known as pathogenesis-related (PR) proteins. The term "PR proteins" refers to a collection of several proteins that are produced both by phytopathogens and signalling molecules involved in plant defence [4]. Salicylic acid (SA) and jasmonic acid (JA) defensive signalling pathways are activated after a pathogen exposure, which further causes the accumulation of PR proteins that reduce pathogen burden or disease start in organs of uninfected plants in figure1 [5]. Additionally, PR proteins are widely dispersed across the plant kingdom, are found in all plant organs, and make up 5–10% of the total proteins in leaves [6]. Despite the fact that PR proteins have been identified numerous times before, their prediction and occurrence in various plant species is still mostly unknown. Since there is a lot of information available regarding PR genes, various research have used for data analysis and bioinformatics based modelling to better understand them [7]. Also numerous studies have focused on predicting and analysing various abiotic stress genes using machine learning, including forecasting activity and predicting mode of action and its relationship among different abiotic stresses [8]. Little research, however, focused on the issue of foretelling biotic stress factors using machine learning approaches. There's a necessity for a comprehensive framework inclusive of detailed elements that demonstrate data analytics and machine learning models specifically tailored for PR proteins associated with biotic stress While majority of genome data analysis of genes have often concentrated on few species of model plants based on amino acid sequence composition, machine learning algorithms have the potential to be used for desired gene prediction within and across species [9].

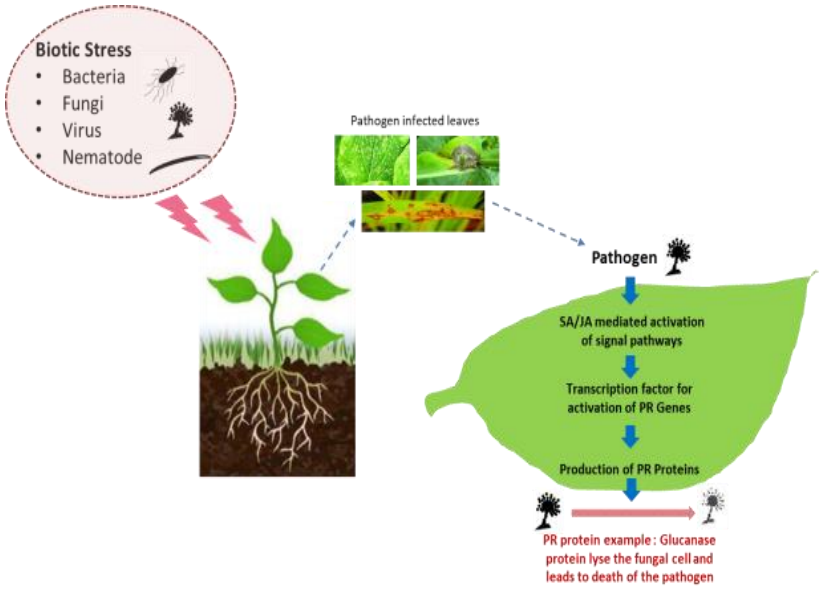


FIGURE 1. Schematic representation of plant response against pathogenic attack with PR-proteins

Models within machine learning, such as support vector machines, decision trees, and Naive Bayes, have demonstrated significant efficacy in predicting protein structures [10]. This study introduces a method that utilizes three distinct machine learning approaches to evaluate the accuracy of predicting HSPs using established models. Our research aims to assist researchers in choosing suitable algorithms when analyzing particular PR proteins from a plant database.

2. MATERIALS AND METHODS

Dataset: For a Biotic stress category, the protein sequences have been taken from uniprot data base for various plants. For the construction of positive and negative datasets 3677 PR proteins were used as positive data, and 2537 PR proteins were used as negative data.

Methodology:

The features have extracted from the various feature selection technique AAC, DC and CTD and it was applied as input data for classification model building. The machine learning models RF, NB and KNN were used to select the best model for PR protein versus non PR protein classification and were used to findthe prediction accuracy and performance measures. Figure 2 illustrates the prediction model procedure.

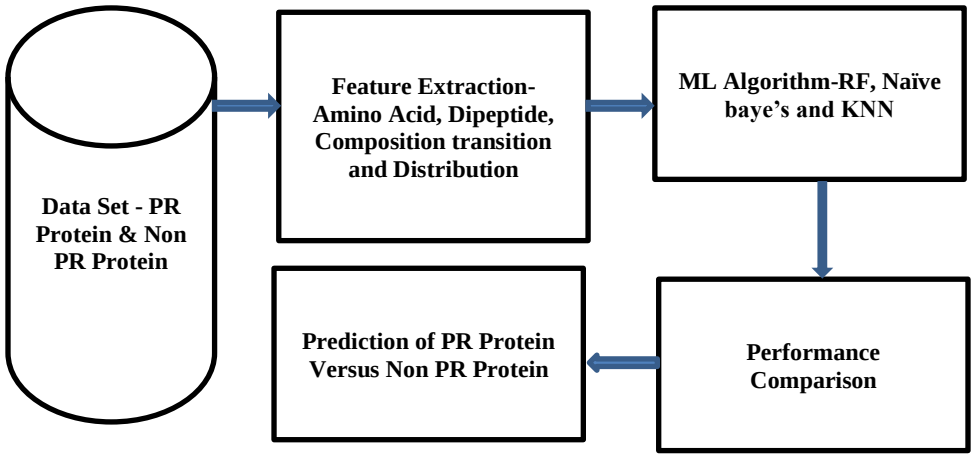


FIGURE 2. Model to predict PR protein sversus non PR proteins

Features Selection Methods:

The prediction of PR protein versus non PR protein is determined by the proper selection of classifier and appropriate set of parameters applied for estimation. Three different feature extraction methods such as Amino Acid Composition (AAC), Dipeptide Composition (DC) and Composition Transition and Distribution (CTD) were computed and used as input data to build the classification models. The python package PyBioMed is used to extract the features of amino acids from the protein dataset. 20 features are extracted from AAC, 400 features from DC and 147 features from CTD. Totally 567 features obtained through this features selection methods to predict the PR.567 features were acquired through this method based on amino acids, dipeptide and physiochemical properties.

1.Amino Acid Composition (AAC): AAC feature selection method plays a pivotal role in characterizing proteins by quantifying the relative frequencies of individual amino acids within a sequence. This method offers crucial insights into protein structure and function, capturing the essence of a protein's composition. By encoding the sequence in terms of its constituent amino acids, AAC provides a fundamental basis for various bioinformatics analyses, such as protein classification, function prediction, and structure determination. A given protein's composition consists of a sequence comprising a set of 20 amino acids.In this approach, the representation of amino acids involves a 20-component vector .The following equations were used to evaluate the property of amino acid character in the protein sequence. A protein sequence represented by 'Protein' and length number 'Number' can be characterized as a sequence y1, y2, y3….y_n, where y_1, y_2,…. $y_n$are the amino acids. The result comprises the existence of every amino acid in a sequence

$$\textbf{AAC of Xi} = \textbf{Quantity of occurrences of Xi in Protein / Number}$$

2. Dipeptide Composition (DC) : The DC feature selection method stands as a beacon of innovation in bioinformatics, wielding its power to filter complex protein sequences into meaningful patterns.This approach employs 400-dimensional vectors derived from a protein sequence (20 x 20 dimensions). These equations were employed to assess the characteristics and behavior of individual amino acids within the specified protein sequence.

$$\textbf{AAC of XiXj} = \textbf{Number of occurrences of XiXj in Protein / Number}$$

For all 1<i,j<=20 .

3.Composition Transition and Distribution (CTD): The feature selection method of CTD within a protein sequence stands as an innovative tool, utilizing its capabilities to elaborate protein structures into recognizable patterns. CTD encompasses various properties of amino acids including hydrophilicity, mass, hydrophobicity, polarity, charge, solvent solubility, secondary, and tertiary structure. These parameters were employed to define distinctive traits for the classification model. In total, 147 descriptors were created specifically for a given protein sequence to facilitate categorization.

A. Machine learning algorithms

Machine learning algorithms revolutionize plant protein classification by binding the power of computational intelligence to decipher complex patterns within protein sequences. The algorithms RF,NB and KNN were used to analyze the vast array of features inherent in plant proteins. By assimilating information from diverse protein characteristics such as amino acid composition and physicochemical properties {11][12]

Random Forest : RF was the one of the most popular ensemble learning methods and has very broad applications in data mining and machine learning. RF particularly used for high-dimensional genomic data analysis. The system RFPDR were developed to predict the disease resistance proteins for Plant [13]

Naïve Bayes : The Naive Bayes algorithm stands as a stalwart in the realm of protein classification, relying on its probabilistic foundation and assumption of feature independence to discern patterns within intricate protein sequences. Operating on the principles of Bayes' theorem, it calculates the probability of a protein sequence belonging to a specific class based on the probabilities of individual amino acids or features. Within eukaryotic organisms, microRNAs (miRNAs) stand as crucial regulatory molecules. Employing a Bayesian approach, our method effectively ranks potential miRNAs, enabling the scoring of those miRNAs that might otherwise be disregarded by alternative methods.[14]

KNN: The KNN algorithm serves as a versatile tool in the realm of plant protein classification, operating on the principle of similarity within a feature space. In the context of plant proteins, KNN navigates the complex landscape of protein sequences, grouping them based on their proximity in feature space. By examining the characteristics and properties of neighbouring proteins, KNN assigns an unclassified plant protein to a class based on the consensus of its nearest neighbors [15].

3. RESULTS AND DISCUSSION

Performance evaluation in machine learning constitutes a vital step, acting as the compass guiding the effectiveness and reliability of deployed models. It encompasses a suite of metrics and techniques designed to quantify and assess how well a machine learning model generalizes to new, unseen data. Through a plethora of evaluation measures such as recall, precision, F-measure (F1) and Accuracy are expressed in the resulting formula and performance evaluation scrutinizes a model's predictive prowess and potential biases.Cross-validation techniques, including K-fold and leave-one-out, play a crucial role in ensuring robustness and reliability by validating model performance across different subsets of data. The Receiver Operating Characteristic (ROC) curve serves as a dynamic graphical representation that elucidates the trade-offs between true positive rate (recall) and false positive rate (1-specificity) across varying classification thresholds. This powerful evaluation tool provides a comprehensive snapshot of a classifier's performance, particularly in binary classification problems.

$$\text{Recall}=\left(\frac{TP}{TP+FN}\right)*100$$

$$\text{F1}=\left(\frac{2*Pr*Sn}{Pr+Sn}\right)*100$$

$$\text{Precision}=\left(\frac{TP}{FP+TP}\right)*100$$

$$\text{Accuracy}=\left(\frac{TP+TN}{TP+FN+TN+FP}\right)*100$$

Predicting with individual features of AAC,DC,CTD

In this research several protein data features used to assess their significance and the measured classification accuracy of each feature using the machine learning algorithms. Based on the individual features AAC , RF algorithm predicts the PR protein with the accuracy of 93% ,95% in DC. In CTD naïve bayes algorithm predicts the PR proteins with 96% accuracy.

TABLE 1. Using the features of AA, DC, CTD

Algorithm	Precision	Recall	F1	Accuracy	ROC	CV
AAC						
RF	93	91	92	90	97	89
Naïve Bayes	82	80	81	77	84	77
KNN	77	89	82	78	86	79
DC						
RF	95	93	94	93	98	92
Naïve Bayes	89	80	84	82	90	85
KNN	81	92	86	82	93	82
CTD						
RF	95	92	93	92	98	91
Naïve Bayes	96	24	39	54	82	52
KNN	70	88	78	70	80	71

Predicting with the combined features of AAC+DC+CTD:

Predicting using the combined features of AAC, DC and CTD represents a holistic approach in predictive function prediction. The synergistic effect of AAC, DC, and CTD empowers machine learning algorithms to discern intricate patterns and relationships within the data, enhancing the predictive capacity and yielding more robust and reliable outcomes in protein versus non Preprotein using the RF algorithm reaches 94%.

TABLE 3. Combined features of AA, DC & CTD

Algorithm	Precision	Recall	F1	Accuracy	ROC	CV
RF	94	99	92	94	99	92
Naïve Bayes	82	90	83	82	90	83
KNN	71	81	71	71	81	71

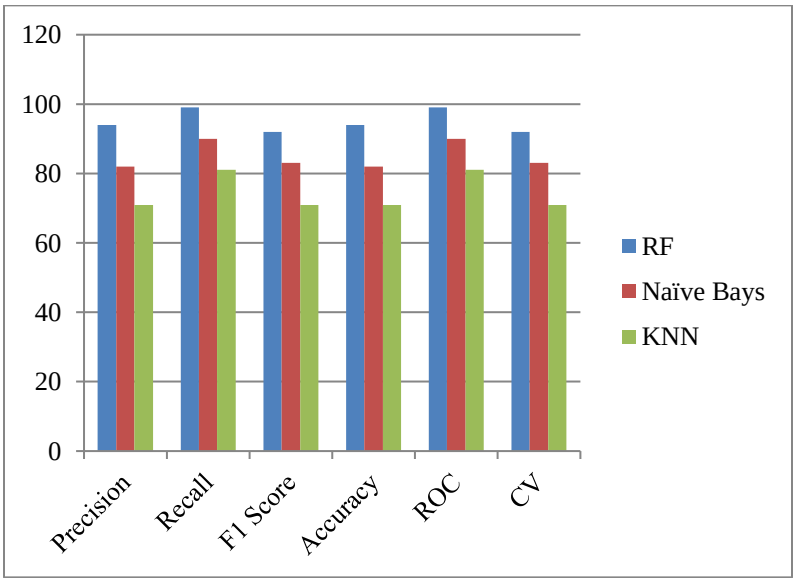


FIGURE 3. Performance metrics with combination of features

4. CONCLUSION

The application of machine learning algorithms for predicting pathogenesis-related proteins in plants represents a significant leap forward in understanding plant defense mechanisms. Through this approach, the system PRED-PATHO demonstrated the potential to revolutionize how to identify and comprehend these crucial proteins, offering a promising avenue for further research and practical applications in agriculture, disease management, and biotechnology. One of the most promising candidates for creating numerous stress-tolerant crop types is PR genes, which have previously been shown to confer improved resistance to both biotic and abiotic stresses. This study aims to employ the supervised algorithm models, including RF, NB, and KNN to identify specific proteins such as plant PR proteins from extensive datasets. The PRED-PATHO system introduces a machine learning methodology designed to predict plant PR protein.As we continue to refine these predictive models, we pave the way for a deeper comprehension of plant-pathogen interactions, fostering resilient crop varieties and enhancing our capacity to safeguard plant health and food security in an evolving environment.

REFERENCES

[1]. Velásquez AC, Castroverde CDM, He SY. Plant-Pathogen Warfare under Changing Climate Conditions. Curr Biol. 2018 May 21;28(10):R619-R634. doi: 10.1016/j.cub.2018.03.054. PMID: 29787730; PMCID: PMC5967643.

[2]. Meemken, Eva-Marie &Qaim, Matin. (2018). Organic Agriculture, Food Security, and the Environment. Annual Review of Resource Economics. 10. 39-63. 10.1146/annurev-resource-100517-023252.

[3]. Kant MR, Jonckheere W, Knegt B, Lemos F, Liu J, Schimmel BC, Villarroel CA, Ataide LM, Dermauw W, Glas JJ, Egas M, Janssen A, Van Leeuwen T, Schuurink RC, Sabelis MW, Alba JM. Mechanisms and ecological consequences of plant defence induction and suppression in herbivore communities. Ann Bot.

2015 Jun;115(7):1015-51. doi: 10.1093/aob/mcv054. PMID: 26019168; PMCID: PMC4648464.

[4]. Jain, Deepti & Khurana, Jitendra. (2018). Role of Pathogenesis-Related (PR) Proteins in Plant Defense Mechanism. 10.1007/978-981-10-7371-7_12.

[5]. Chaturvedi, R., Shah, J. (2007). Salicylic Acid in Plant Disease Resistance. In: Hayat, S., Ahmad, A. (eds) Salicylic Acid: A Plant Hormone. Springer, Dordrecht. https://doi.org/10.1007/1-4020-5184-0_12

[6]. Sajad Ali, Bashir Ahmad Ganai, Azra N Kamili, Ajaz Ali Bhat, Zahoor Ahmad Mir, Javaid Akhter Bhat, Anshika Tyagi, Sheikh Tajamul Islam, Muntazir Mushtaq, Prashant Yadav, Sandhya Rawat, Anita Grover, Pathogenesis-related proteins and peptides as promising tools for engineering plants with multiple stress tolerance,MicrobiologicalResearch,Volumes 212–213,2018,Pages 29-37,ISSN 0944-5013,https://doi.org/10.1016/j.micres.2018.04.008.

[7]. Bayat A. Science, medicine, and the future: Bioinformatics. BMJ. 2002 Apr 27;324(7344):1018-22. doi: 10.1136/bmj.324.7344.1018. PMID: 11976246; PMCID: PMC1122955.

[8]. Meher, P., Begam, S., Sahu, T., Gupta, A., Kumar, A., Kumar, U., Rao, A., Singh, K. and Dhankher, O., 2022. ASRmiRNA: Abiotic Stress-Responsive miRNA Prediction in Plants by Using Machine Learning Algorithms with Pseudo K-Tuple Nucleotide Compositional Features. International Journal of Molecular Sciences, 23(3), p.1612.

[9]. Libbrecht, M. and Noble, W., 2015. Machine learning applications in genetics and genomics. *Nature Reviews Genetics*, 16(6), pp.321-332.

[10]. Rácz A, Bajusz D, Miranda-Quintana RA, Héberger K. Machine learning models for classification tasks related to drug safety. Mol Divers. 2021 Aug;25(3):1409-1424. doi: 10.1007/s11030-021-10239-x

[11]. Cai YD, Liu XJ, Xu XB, Zhou GP. Support vector machines for predicting protein structural class. BMC Bioinformatics. 2001; 2(1):1.

[12]. Chaurasiya M, Chandulal GB, Misra K, Chaurasiya VK. Nearest-neighbor classifier as a tool for classification of protein families. Bioinformation. 2010; 4(9):396-398.

[13]. Simón D, Borsani O, Filippi CV. RFPDR: a random forest approach for plant disease resistance protein prediction. PeerJ. 2022 Apr 22;10:e11683. doi: 10.7717/peerj.11683. PMID: 35480565; PMCID: PMC9037127.

[14]. Douglass, S., Hsu, S.-W., Cokus, S., Goldberg, R.B., Harada, J.J. and Pellegrini, M. (2016), A naïve Bayesian classifier for identifying plant microRNAs. Plant J, 86: 481-492. https://doi.org/10.1111/tpj.13180

[15]. Morgan, Maria & Blank, Carla &Seetan, Raed. (2021). Plant disease prediction using classification algorithms. IAES International Journal of Artificial Intelligence (IJ-AI). 10. 257. 10.11591/ijai.v10.i1.pp257-264.