Eng. Sci., 2022, **19**, 205–214

Engineered Science

DOI: https://dx.doi.org/10.30919/es8d712

Areca Nut Disease Dataset Creation and Validation using Machine Learning Techniques based on Weather Parameters

Rajashree Krishna,[1,*] Prema K V[2] and Rajat Gaonkar[3]

Abstract

Areca nut crop yield is affected by many diseases caused due to heavy rainfall and high relative humidity. An early prediction of crop disease based on weather data can help farmers to take preventive measures. Many machine learning applications are deployed to detect the disease through image data. The proposed study is the first approach for creating a novel dataset and developing the weather-based areca nut disease prediction model. Historical weather data *i.e*. temperature, rainfall, relative humidity, sunshine, wind direction, and wind speed are collected from the Udupi weather station. Fruit rot disease data are collected through farmer surveys, disease management recommendations, and research literature. These data are integrated and correlated to create the final dataset which is validated and compared using a statistical method, decision tree regression (DTR), multilayer perceptron regression (MLPR), random forest regression (RFR), and support vector regression (SVR) models. Principle component analysis, branch and bound, and wrapper feature selection techniques are used to select the weather parameters contributing to more accurate prediction. The observation shows that RFR gives 0.9 mean absolute error (MAE) as the lowest value among many models and SVR gives 1.7 MAE as the highest error after feature selection.

Keywords: Machine learning; Areca nut; Crop disease; Prediction; Feature selection; Weather parameters.
Received: 16 December 2021; Revised: 06 May 2022; Accepted: 25 May 2022.
Article type: Research article.

1. Introduction

India ranks first in areca nut production globally and many farmers depend on it for their livelihoods. Areca nut or betel nut is a plantation crop distributed over Southeast Asian countries including India, Malaysia, China, *etc*. Areca nut is mainly used in mastication with betel leaves by 600 million people daily and is also used in religious and social ceremonies.[1] Nearly 853,000 metric tons of areca nut are produced in 518,000 Hectares every year.[2] Areca nut production in India is more than 50% of the world's areca nut production, and states like Karnataka, Kerala, Goa, Meghalaya, Assam, and West Bengal are the principal producers.[3] According to Hindu customs, the areca nut is a unique worshipful element; without this, worship will not be successful.[4] Like other crops, the areca nut is also vulnerable to disease occurrence by insect pests and environmental factors throughout the year at every stage of its life. Areca nut suffers from diseases like fruit rot, foot rot, yellow leaf disease, bud rot, *etc*.

Nowadays usage of chemical fertilizers for areca nut crops has reduced due to a lack of human resources, which increases the development of fruit rot disease. The fruit rot disease is caused by Phytophthora spp fungus and it results in partial or total crop loss in individual palms or death of the palm itself. Usually, fruit rot disease outbreaks after 15 to 20 days of monsoon onset and continues up to the rainy season's end. Pre-infection treatment is better than post-infection treatment in terms of fungicide resistance suppression and control cost.[5] Therefore, it is indispensable to prevent fruit rot disease, and predicting disease incidence at an early stage will be very supportive in achieving the above-mentioned aim.

As the population increased, pressure developed in the agricultural system to satisfy the demands, which initiated digital agriculture. Machine learning (ML) and deep learning (DL) play a very important role in the agriculture domain, also the advancements in ML can be observed in agricultural tasks.[6] Applications of ML in the crop section include yield prediction, crop quality recognition, disease detection, species

[1] *Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, 576104, Karnataka, India.*
[2] *Department of Computer Science and Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Bangalore, 560064, Karnataka, India.*
[3] *Department of Electronics and Communication Engineering, PES University, Bangalore, 560085, Karnataka, India.*
**Email:* raji.krish@manipal.edu (R. Krishna)

recognition, and forecasting plant and crop diseases.[7,8] Different machine learning algorithms like support vector machines, decision tree algorithms, artificial neural networks, and random forest classification algorithms are used to develop crop disease, forecasting models.[8] But ML models have been used very less in weather-based areca nut disease prediction.

There are two categories of crop disease prediction models based on the processing of input parameters: that are image-based models and weather-based models.[9–11] Prediction using images has the limitation of predicting the disease once the disease is in incidence. So the crop grower cannot prevent the disease from occurring. But the prediction based on weather parameters will predict the disease before incidence, which will help the crop grower to take preventive action. Hence this research will create the dataset for areca nut crop disease and predict the disease incidence using statistical and machine learning models based on climate data.

Many investigations have been done previously on crop disease prediction and identification. Vinod Kanan and Sanjeev Kumar recently forecasted areca nuts' fruit rot disease using IoT and machine learning by considering weather parameters.[12] Support vector regression and random forest classifier methods are used for prediction. Some sensors are used to capture the weather data, and soil humidity and they found promising results.

In the previous study, Sujatha *et al*. observed the influence of climate and weather parameters on the areca nut and cocoa crop yield.[13] Heavy rainfall and large-intensity rains will spread the fruit rot disease. The Bordeaux mixture is sprayed on the areca nut bunches once in 45 days, to prevent the disease from occurring. Ajith Danti and Suresha M classified areca nuts based on their texture.[14] They used decision tree classification algorithms which take images as input. Using computer vision techniques the same authors have also classified and segmented raw areca nuts.[15] A three-sigma control unit is used to set the image's background color. Here the binary classification is used for two classes called boiling nuts and non-boiling nuts. Mariusz Wrzesie´n *et. al*.[16] predicted Apple scab using a random forest algorithm. Apple scab is a pest that occurs when leaves are wet for a long time at a given temperature.

Hyo-suk Kim *et al*.[5] used predicted weather data produced by the Unified Model and observed weather data by automated weather stations to forecast rice crops' bacterial grain rot disease. They found that the major advantage of using predicted weather data is that the information about disease forecasts should be available before genuine infection by a pathogen, allowing farmers to take proper disease management plans. Regardless of weather data sources, the model could predict the disease warning. Aman Sharma *et al*.[17] experimented to find out the relationship between the incidence of bacterial blight disease in cotton with weather parameters. Correlation result reveals that temperature range, relative humidity, and long duration of bright sunshine were highly favorable for the disease development. Lucas Eduardo de Oliveira Aparecido *et al*.[18] used machine learning models like K neighbors regressor, random forest regressor, multiple linear regressors, and basic neural networks, for forecasting the incidence of coffee pests and diseases. They used field and weather data obtained from the coffee plantation in Brazil as independent parameters for the model. In southern Idaho, potato crop disease called Late blight is forecasted using logistic regression analysis.[19] The paper's objective is to check whether the late blight disease occurrence depends on whether parameters and the model's capacity to predict the disease incidence in other regions.

BLITE-SVR is a support vector regression (SVR) model used to forecast potato late blight disease.[20] This disease occurs only when specific weather conditions like 85% and above humidity and total rainfall of 30mm continue for seven days. The author calculated the correlation between the first date of disease occurrence and the different weather parameters. This correlation value is used in the equation to calculate the threshold SVR score. If the score value exceeds the threshold value and continues for six days, then the model predicts the actual date of disease occurrence. The authors used machine learning techniques to consider potato late blight disease forecasting case study.[21,22] They used an inner mathematical model called SimCast, which takes temperature and humidity as inputs and provides blight units as outputs. These blight units (BU) are called risk indexes, their value varies from 0 to 7 and it is used to calculate the severity of late blight infection. Daily BU, humidity, and the daily temperature are given as input to support vector machine and artificial neural network model. Potato late blight disease prediction is also done by ANN model with the help of weather parameters and achieved 90.9% accuracy.[23]

M. Karagiannopoulos *et al*.[24] summarized the different feature selection methods used for regression problems. They have taken the 12 different datasets for the experiment purpose. Five different types of wrapper feature selection techniques like forwarding selection, backward selection, best first backward selection, best first forward selection, and genetic search selection are used and verified through regression techniques. The author says that wrapper methods give better results when compared with filter methods. Naoual *et al*.[25] have reviewed the wrapper feature selection methods. It is always difficult to rank the feature selection method because it varies from one feature set to another feature set. Nevertheless, feature selection is very crucial in machine learning to get good performance in prediction/classification.
It has been found that there are forecasting models developed for crops like potato, wheat, rice, mango, coffee, grape, maize, mustard, *etc*.[11,18,19,21,23,26,27] As per the author's knowledge there is no prediction model developed for areca nut crops, which is probably due to the unavailability of the dataset in the public domain.[28] Hence the study focuses on creating a novel areca nut disease data set and validating the same using statistical method and supervised machine learning algorithms. Multiple

regression (MR), SVR, RFR, DTR, and basic neural network regression technique MLPR are executed on the created dataset, and results are compared. The remaining content of this paper is organized as follows. Section 2 describes dataset preparation, machine learning techniques, and experimental procedure. In section 3, the results are discussed and section 4 gives the conclusion and future enhancements.

2. Experimental section

The primary purpose of the present study is to prepare and validate the dataset which combines the areca nut disease data and historical weather data. This work is the first attempt to determine the relationship between the fruit rot disease incidence in areca nut based on environmental factors and also to validate the effectiveness of the machine learning models on this relationship. Fig. 1 shows the data flow diagram for creating and validating the dataset by predicting the disease incidence severity in the areca nut.

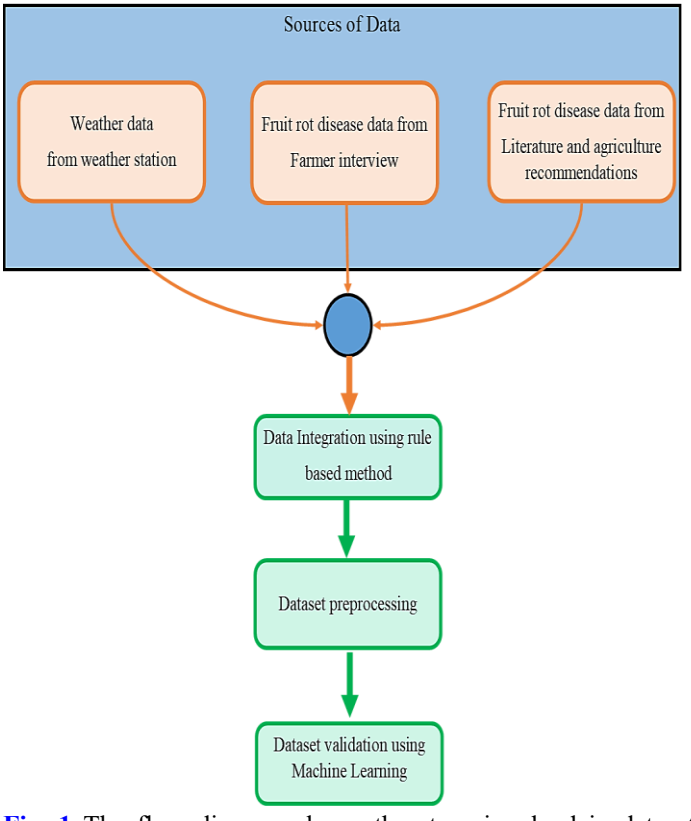

Fig. 1 The flow diagram shows the steps involved in dataset creation and validation.

2.1 Data set preparation

The region used for the study is Udupi, the city in the Indian state of Karnataka with a latitude of 13.3409° N and longitude of 74.7421° E. Areca nut disease data readings for this region are not available in the public domain. Therefore, it is essential to create the dataset with the help of weather data and disease data collected from the Udupi weather station, farmer surveys, disease management recommendations, and research literature.[2,4]

2.1.1 Historical climatic data

Zone Agricultural and Horticultural Research Station, Brahmavar, Udupi provides the historical weather data for 21 years from 2000 to 2020 required for experimentation. The fruit rot disease occurs during the monsoon season between June and October. So the weather data from June to October is considered for the dataset. The overall parameters present in the dataset are Minimum Temperature (MinT), Maximum temperature (MaxT), Rainfall (RF), Relative Humidity (Rh), Clouds (amount of cloud cover), Sunshine hours (SS), Wind speed, Target/Score value to measure disease incidence. Table S1 tabulates daily weather parameters for June 2020 as sample weather data.

2.1.2 Areca nut disease score data

Arecanut disease incidence recordings for any region are not available in the public repository. So the information related to disease data is taken from Thotagarika Ilaake Doddanagudde and disease management recommendations. The information is also collected by interviewing nearly 50 farmers in and around Udupi. Integrating weather data and areca nut disease data plays a significant role in data set creation. Information collected from farmers, agriculture departments in Udupi, and literature are collectively used in the dataset creation process. If there is heavy rainfall continuously for 15 to 20 days, then there is a chance that fungus will generate. Only rainfall is insufficient, but a humidity of more than 90 and a low temperature of 20 to 23 degrees are very supportive of increasing the fungus growth and hence disease outbreaks.[29] The collected official proofs say that in Udupi there was a disease outbreak in 2013, and 2018 but the farmer's survey says that there was a disease outbreak in 2019 and 2020 as well.

Since the areca nut disease recordings are not available in the public domain, the rule-based classification method is used to connect weather data and areca nut disease data to predict disease incidence.[30] In the previous study, the rule-based algorithm is used to build a weather-based dataset for rice blast disease. The disease severity index is considered output, a numerical value between 1 to 6.[31] In this method, an IF-THEN rule is an expression of the form LHS ⇒ RHS where LHS is a condition that must be met to derive the conclusion shown in RHS. Accordingly, in the present study, the areca nut disease score is a numerical value between 1 to 35 is calculated with the help of the following rules.

1. If the rainfall is more than 15 mm, the temperature is less than 24 °C and the humidity is more than 90% the score value will be incremented by one until it reaches 35 as a threshold value.

2. The score value also increases if there is intermittent rainfall and sunshine; that is if rainfall is more than 5 mm and sunshine is more than 5 hours, the score value increases.[2]

3. If the rainfall reduces and reaches less than 10 mm, and the temperature increases more than 24 °C then the score value will be reduced by one.

Therefore, the fruit rot disease severity was categorized using measures of 0-35 score value. Table 1 displays the meaning of each class.

Table 1. Different types of classes and their meaning.

Classes	Meaning
Class 0	Resistance
Class 1	Moderate Resistance
Class 2	Susceptibility
Class 3	Susceptibility to severe

Class 0 considers the score ratings of 0-15 as resistance, the areca nut crop will not show any fungal activity, if the weather pattern continues then the score rating of 15-25 indicates class 1 as moderate resistance. At this stage, the fungus production starts and initial symptoms appear as yellowish water-soaked lesions on the nut surface. In susceptibility class the score ratings will be 25-34, slowly the lesions will spread to the whole nut and immature nuts will start shedding. When the score rating becomes 35, the disease becomes severe and it occupies the whole plantation area, it can be considered as class 3 according to the classification strategy.

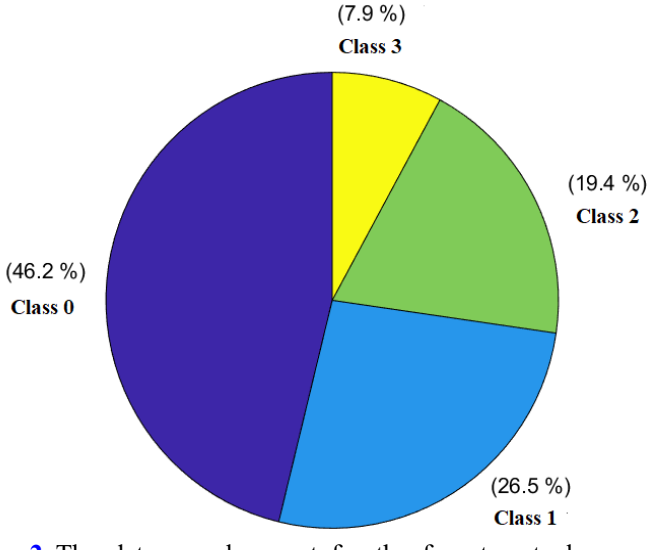

Fig. 2 The data sample count for the four target classes as mentioned in Table 1.

The integrated data set contains 3152 instances, 11 features, and 1 target value (score rating). Fig. 2 shows the data sample distribution over different classes. When 21 years of data is considered, among 3152 samples, 46.2% of samples come under the class 0 category, and 26.5% of samples come under class 1. Similarly, 19.4% of samples are under class 2, and class 3 contains very less samples which are 7.9% only. Therefore, the present study experiments with the imbalanced dataset.

The graph in Fig. S1 indicates that the areca nut crop's disease incidence severity was at its peak in 2001, 2003, 2011-2013, and 2015-2020. In 2009 disease was not found in areca nut crops in the Udupi region.

After integrating the weather data and disease data, it has been checked for prediction accuracy to validate the dataset using MR, SVR, RFR, DTR, and MLPR.

2.1.3 Data preprocessing

This experiment uses daily-based weather data as time series data. Preprocessing the time series data for supervised machine learning has a significant influence on prediction accuracy; therefore, it is essential in the model. The date column in the dataset is set as an index. Missing values are filled with the mean value. The incidence of the disease is always depending on the minimum 15 days weather pattern. So the input data variables are shifted by 15 days. Hence the time series data is preprocessed so that the newly created data set will fit the supervised machine learning regression models. Section 2.2 details the different regression models applied to validate the dataset.

2.2 Data set validation models

The most commonly applied machine learning regression models and statistical models are used for the first time to predict the areca nuts' fruit rot disease incidence. Fig. 3 shows the list of techniques used in the experiment in different stages. The data set is divided into the training set and testing sets with a 70:30 ratio. Weather parameters are considered independent variables and disease score value is considered a dependent variable. The dataset proposed in this study evaluates data from the past 21 years. Data from 16 years are used for the model training and data from 5 years is used for model testing.

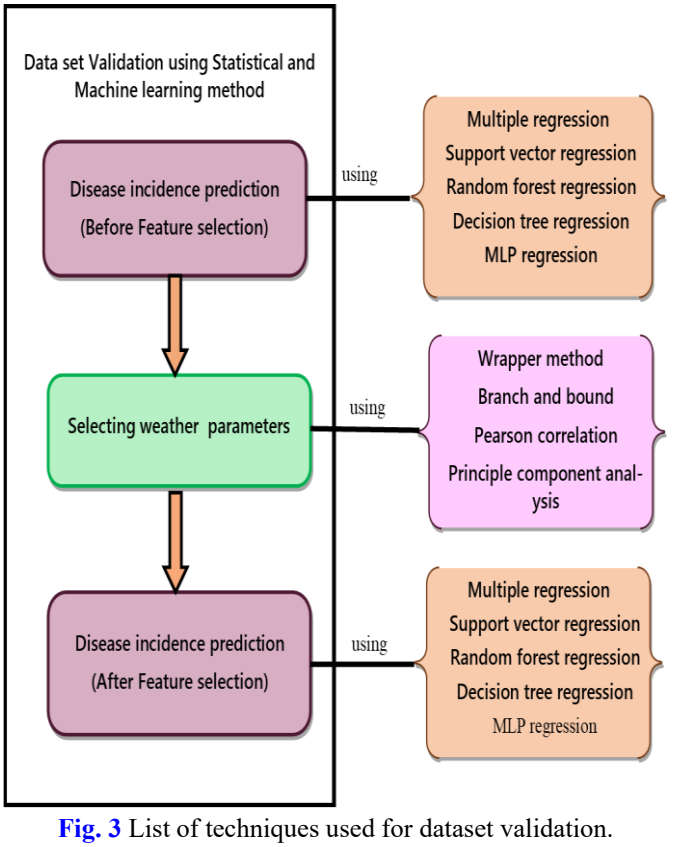

Fig. 3 List of techniques used for dataset validation.

MR, SVR, DTR, RFR, and MLPR techniques are operated to predict the crop disease score value before selecting weather features. Branch and bound, wrapper method, principal component analysis, and Pearson correlation techniques determine the effective weather parameters. After discarding some weather features, again the MR, SVR, DTR, RFR, and MLPR techniques are used to predict the disease score ratings.

2.2.1 Statistical method

MR is a statistical method used to know the impact of several independent variables on a single dependent variable. In the present study, the MR is used to find the relationship between the weather parameters and fruit rot disease in the areca nut. The MR model is expressed as shown in Equation (1):

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_k x_k + \varepsilon \quad (1)$$

where, y is the disease score value as the dependent variable, x_1, x_2, x_3 are weather parameters as independent variables, β_1, β_2, β_3 are feature weights, k represents the number of features and ε is the residual term of the model.

2.2.2 Support vector regression

SVR is similar to the support vector machine with a hyperplane and boundary margin. Recently the SVR has been used in travel time prediction, electricity price forecasting, financial market forecasting, estimation of power consumption, *etc*.[32,33] Other regression models will minimize the error between the actual value and predicted value but the SVR aims to fit the hyperplane within the predefined error value ε as shown in Fig. 4.

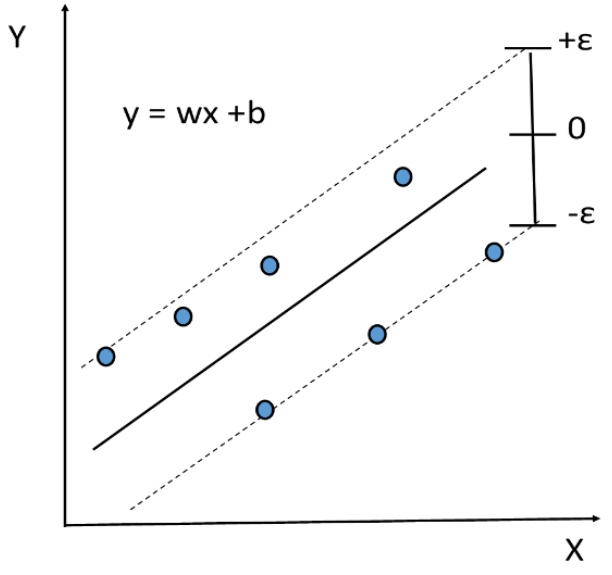

Fig. 4 Support vector machine – Regression.[30]

The constraints to be used to define the error value ε are given in Equation (2).

$$y_i - wx_i - b \leq \varepsilon \quad and \quad wx_i + b - y_i \leq \varepsilon \quad (2)$$

where w is a vector normal to the hyperplane, b is an offset and x is a sample vector. SVR model from the Scikit-learn library is used along with RBF as a kernel function. The other SVR parameters used for the proposed study are epsilon = 0.1, Kernel coefficient gamma = scale, and regularization parameter C = 1.0.

2.2.3 Decision Tree Regression

DTR is a tree-based prediction algorithm. It is used to predict the numeric value of the dependent variable. Decision tree regression is used in software fault prediction. The conventional decision tree technique builds the tree with the root node and leaf node.[34] But the standard deviation reduction method is used in place of information gain to select the different nodes.

Steps used in DTR:

1. The standard deviation of the target value is calculated, *i.e*. S(T).
2. The standard deviation of each branch in each attribute is calculated, *i.e*. S(T, X) is shown in Equation (3).

$$S(T,X) = \sum_{c \in X} P(c) S(c) \quad (3)$$

here, P(c) is the probability of class c, and S(c) is the standard deviation of class c, where c belongs to the predictor.

3. As shown in Equation (4), both the results are subtracted, and it is the standard deviation reduction (SDR) value.

$$SDR(T,X) = S(T) - S(T,X) \quad (4)$$

Standard deviation reduction decreases the standard deviation after a dataset is split into attributes as a decision node. The attribute with the largest SDR is selected as the root node in the tree. The process continues recursively. The regression performance is measured through mean absolute error (MAE), root mean square error (RMSE), and R square.

The error calculation is as follows:

Prediction error = real output - observed output
result = summation (square of prediction error)
MSE = result/number of samples
RMSE = square toot of MSE
MAE = summation (prediction error) / number of samples.

The parameters used to do the experiment are:

(1) Splitter = 'best'; a strategy used to split attributes at each node.
(2) Criterion = 'squared error' used to measure the quality of the split.
(3) Scikit-learn default parameters of the DecisionTreeRegressor() method.

2.2.4 Artificial neural network (ANN)

Over the last few years, ANN has become a prevalent tool to solve the classification and prediction problems in multiple domains but the ANN application in the agriculture domain is significantly less compared with other disciplines.[35] It automatically allows for a non-linear relationship between response variables and predictor variables because of the flexibility in choosing the activation function, the number of nodes, and connections. The authors proved that the ANN is a more efficient prediction model.[27] Fig. 5 shows the block diagram of an artificial neural network with different layers.

The present study considers the following parameters to generate accurate results:

(1) The number of hidden layers is 3 and 64, 32, and 16 neurons are taken in each hidden layer, respectively.
(2) ReLu activation function is used with a learning rate of 0.001.
(3) Adam optimizer is used to optimize the weights.

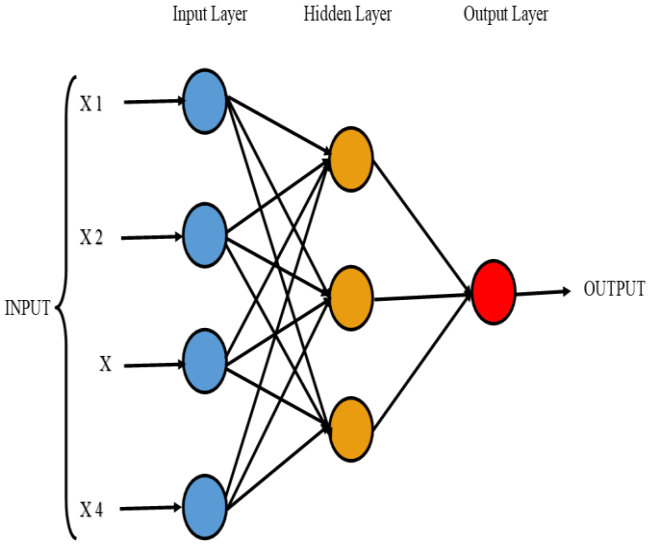

Fig. 5 Multi-layer perceptron with three layers input, output, and hidden.

2.2.5 Random forest regression

RFR is the supervised ensemble learning algorithm. Here prediction result will be given by multiple trees and the average result is considered the final prediction. It is more accurate when compared with a single model. The schematic representation of the random forest is given in Fig. S2. 100 trees in the forest are considered to validate the newly created dataset along with the default parameter values from the scikit learn library.

3. Results and discussion

The present study demonstrates the creation of a weather-based fruit rot disease dataset for areca nut crops. The dataset was then used to develop an early prediction model for areca nuts' fruit rot disease incidence based on weather parameters using statistical and machine learning techniques. Target values/score values are predicted with the help of environmental factors. The prediction accuracy of different models is compared and shown in Table 2.

The performance of the models is assessed by MSE and MAE metrics. From Table 2 it can be observed that MR and RFR have a petite MAE of 0.9 whereas multi-layer perceptron regression has a large MAE of 1.9, and an MSE of 2.0 is provided by the RFR model which is the least value among all models. The actual value and predicted value graph for DTR, SVR, RFR, and MLPR machine learning algorithms are shown in Fig. 6.

Table 2. Prediction accuracy of the different learning models based on MSE and MAE before selecting the features.

Algorithms	MAE	MSE
Multiple regression	0.9	2.1
Support vector regression	1.8	8.0
Random forest regression	0.9	2.0
Decision tree regression	1.2	3.7
Multi-layer perceptron regression	1.9	7.9

The graph shows that the predicted value is almost similar to the real value in all regression algorithms when all the weather features are considered. Among 945 testing samples, only 100 predicted values are displayed.

However, feature selection techniques are used to select the essential features to reduce the error rate. It is observed from the literature that the wrapper methods are suitable to select the relevant features.[24,25,36] And also Principle component analysis (PCA) technique is widely used in the feature selection process. Similarly, branch and bound is a tree-structured feature selection algorithm typically used in supervised machine learning.[37] Pearson correlation is used to confirm the selected features, and the correlation heat map is shown in Fig. 7. This heat map indicates the correlation between the weather parameters and the target value.

The heat map shows that rainfall, cloud cover, and relative humidity are positively correlated with the trigger/target value; temperature and wind direction features are negatively correlated with the target value. Finally, the impact of various feature selection methods on areca nuts' disease incidence prediction is analyzed in the study.

After the conclusion of all feature selection methods, minimum and maximum temperature, rainfall, relative humidity, and sunshine features are selected as independent variables for the prediction process. Wind speed and cloud-based features are removed. Statistical and regression models are applied to the chosen features to observe the error rate and accuracy. After feature selection, the MAE and MSE of different prediction models are compared and shown in Table 3. The RFR model gives the highest accuracy in terms of MAE and MSE compared with other models. The observation also says that the machine learning method is more efficient than the traditional statistical method.

Table 3. Prediction accuracy of the different prediction models based on MSE and MAE after selecting the features.

Algorithms	MAE	MSE
Multiple regression	0.9	2.0
Support vector regression	1.7	6.1
Random forest regression	0.9	1.9
Decision tree regression	1.2	3.4
Multi-layer perceptron regression	1.2	3.3

The actual value and predicted value graph for DTR, SVR, RFR, and MLPR machine learning algorithms after the feature

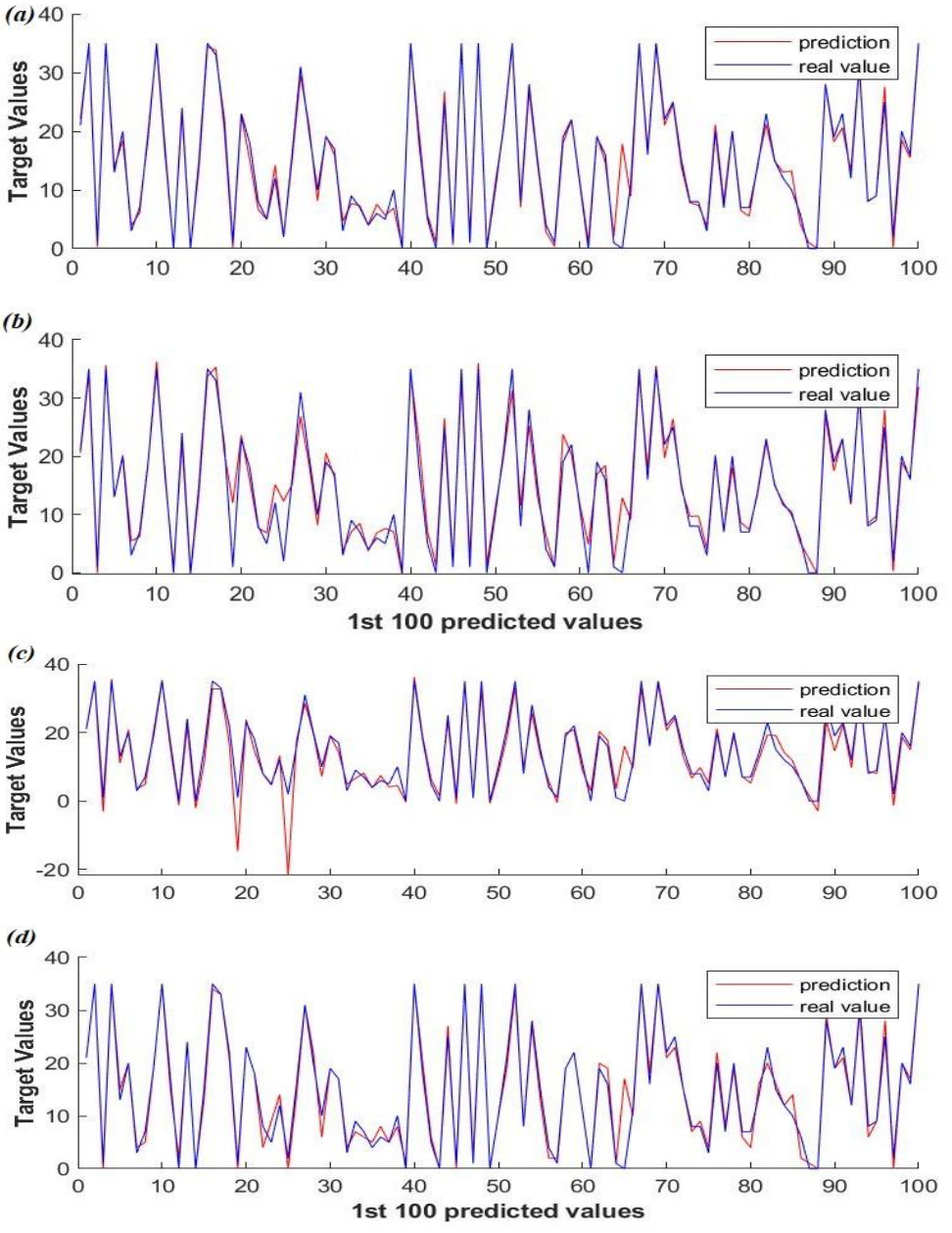

Fig. 6 Prediction graph for different regression models before feature selection: (a) Random Forest Regression (b) Support Vector Regression (c) Multi-layered Perceptron Regression (d) Decision Tree Regression.

	RF (mm)	Max.Temp	Min.Temp	RH-I	RH-II	Cloud I	Cloud II	Wind Dn-I	Wind Dn-II	Trigger
RF (mm)	1.000000	-0.375220	0.032686	0.300235	0.367929	0.401663	0.392594	-0.034552	-0.058042	0.031189
Max.Temp	-0.375220	1.000000	0.041747	-0.259002	-0.417581	-0.418332	-0.419398	-0.002102	0.091616	-0.242843
Min.Temp	0.032686	0.041747	1.000000	0.001769	-0.019272	-0.058224	-0.073980	0.095128	0.001739	-0.133200
RH-I	0.300235	-0.259002	0.001769	1.000000	0.354316	0.268554	0.220732	0.005793	-0.022871	0.203651
RH-II	0.367929	-0.417581	-0.019272	0.354316	1.000000	0.402089	0.466699	-0.055347	-0.091099	0.234897
Cloud I	0.401663	-0.418332	-0.058224	0.268554	0.402089	1.000000	0.657327	-0.052288	-0.081445	0.080554
Cloud II	0.392594	-0.419398	-0.073980	0.220732	0.466699	0.657327	1.000000	-0.091521	-0.086933	0.098205
Wind Dn-I	-0.034552	-0.002102	0.095128	0.005793	-0.055347	-0.052288	-0.091521	1.000000	0.227152	-0.004362
Wind Dn-II	-0.058042	0.091616	0.001739	-0.022871	-0.091099	-0.081445	-0.086933	0.227152	1.000000	-0.005361
Trigger	0.031189	-0.242843	-0.133200	0.203651	0.234897	0.080554	0.098205	-0.004362	-0.005361	1.000000

Fig. 7 Pearson Correlation matrix heat map.

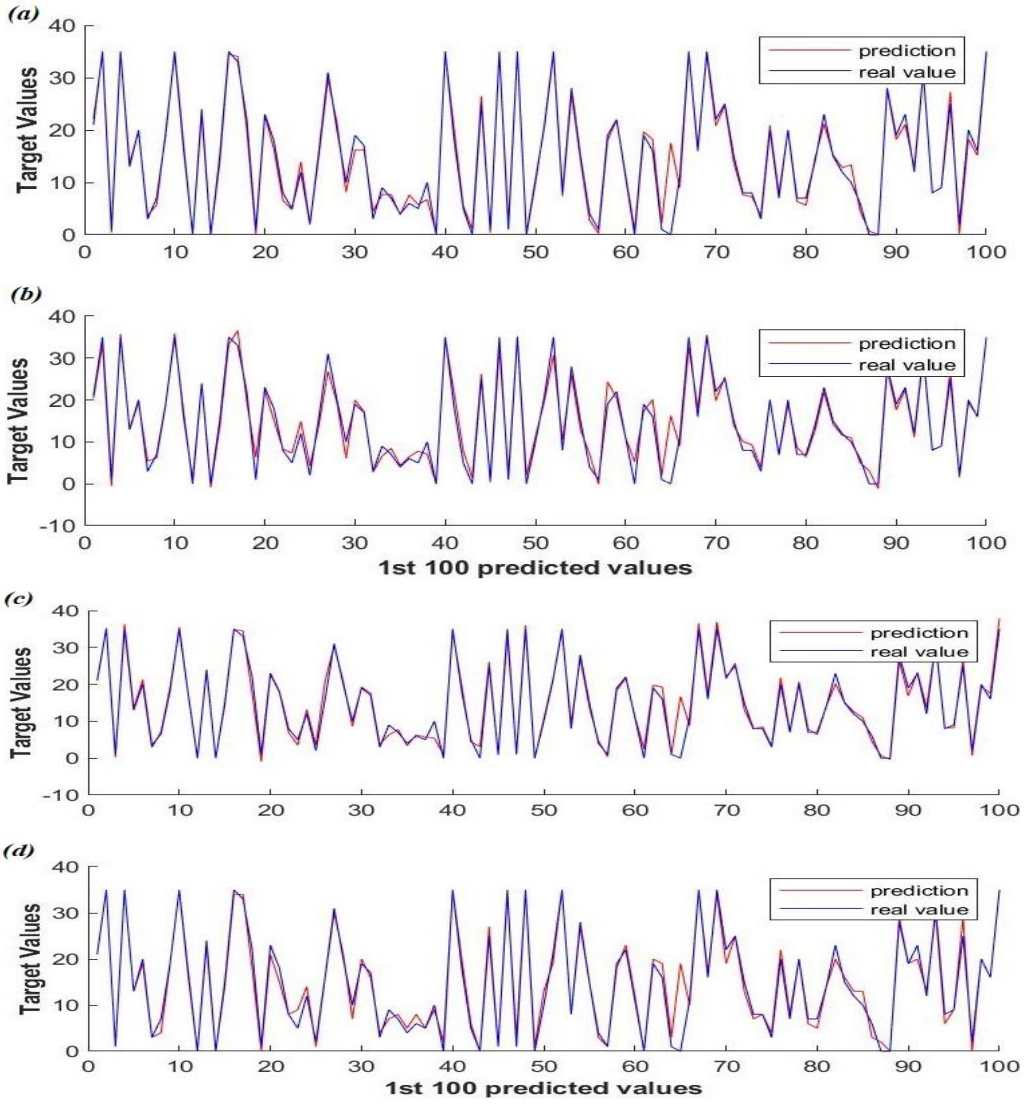

Fig. 8 Prediction graph for different regression models after feature selection: (a) Random Forest Regression (b) Support Vector Regression (c) Multi-layered Perceptron Regression (d) Decision Tree Regression.

selection is shown in Fig. 8. The graph shows that in all regression algorithms when selected weather features are taken into account, the predicted value is almost the same as the real value. It shows only 100 predicted values from 945 testing samples. The collective comparison of MAE and MSE performance metrics before and after feature selection is shown in Fig. 9. According to the graph, the highest prediction performance is given by random forest regression as it uses the ensemble learning method for regression. Followed by decision tree regression, for which the mean absolute error is 1.2. SVR and MLP regression show more error rates comparatively.

In addition, feature selection techniques helped to reduce the error rate measurably; instead of 10 features, only 7 features are used to test the model accuracy. After feature selection, the MLP regression model reduced the MSE from 7.9 to 3.3 followed by SVR decreased its mean square error from 8 to 6.1. Similarly, in SVR mean absolute error is reduced from 1.8 to 1.7 and in the MLP regression model, the reduction in mean absolute error is from 1.9 to 1.2. But there is no accuracy improvement in RFR and DTR after feature selection. Finally, the dataset is validated through the help of statistical and machine learning regression models. Since the Udupi region has many areca nut growers and suffers from yield loss due to disease, it has spurred various research groups to develop the early prediction model. The present study is the first attempt to predict the areca nuts' disease incidence score value. It uses both statistical and machine learning methods for prediction when compared with the previous survey which uses only statistical methods for forecasting late blight on potatoes.[20] Hyo-suk Kim *et al*.[5] used a unified model and automated weather stations (AWS) observed data are used to estimate the grain rot risk in rice, whereas this model uses only AWS weather data. Vinod Kanan and Sanjeev Kumar[12] used only two years of weather and disease data to develop the areca nut disease prediction model and uses only the support vector machine and random forest algorithms. In this research 20 years of weather and disease data are considered and

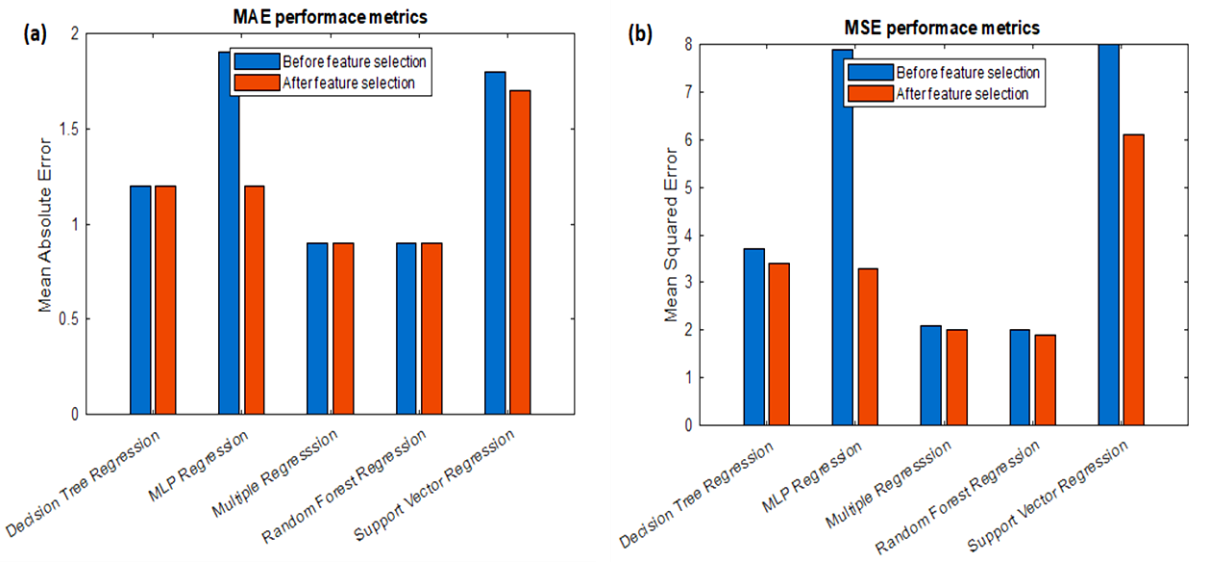

Fig. 9 Comparison of models before and after feature selection: (a) Mean absolute error performance metric and (b) Mean square error performance metric.

applied MR, MLP regression, DTR, SVR, and RFR algorithms. Compared with other research, this study gave preferences to the areca nut crop for which the prediction model is not yet developed, since it helps the areca nut farmers in the Udupi district enormously.

4. Conclusions

Research on areca nut has been going on for 25 years on its usage, side effects, growth, yield, and disease classification. Similarly, many crop disease prediction models using images are available but the weather-based areca nut crop's disease prediction model is not developed due to the lack of a dataset. Hence this is the first attempt to create a dataset and forecast the disease outbreak with the help of weather parameters. The dataset used to predict the disease incidence is prepared by collecting the weather and disease data from Udupi local region. Collected data is assembled through the medium of a rule-based classification method. The dataset is validated by utilizing SVR, DTR, RFR, and MLP regression models. It is found that RFR gives very good accuracy and minimum error rate in the prediction of areca nut fruit rot disease incidence. Nonetheless, different feature selection techniques are used to remove unwanted weather parameters and increase accuracy. It is observed that the removal of features has affected the error rate, which can be further improved. As a future enhancement, the data set size can be increased and can include other features. The foundation used in this study can be broadened to develop a prediction system for the remaining regions. Also, artificial neural networks and deep learning techniques can be applied in the future to enhance prediction accuracy.

Acknowledgments

Thotagarika Ilaake Doddanagudde, Udupi and Zone Agricultural and Horticultural Research Station, Brahmavar, Udupi supports this work.

Conflict of Interest

The authors declare no conflict of interest.

Supporting information

Applicable.

References

[1] V. Paramesh, V. Arunachalam, A. Nikkhah, B. Das, S. Ghnimi, *Journal of Cleaner Production*, 2018, **203**, 674-684, doi: 10.1016/j.jclepro.2018.08.263.

[2] P. Balanagouda, H. Vinayaka, H. P. Maheswarappa, H. Narayanaswamy, *Indian Phytopathology*, 2021, **74**, 561-572, doi: 10.1007/s42360-021-00382-8.

[3] P. S. Deshmukh, P. G. Patil, P. U. Shahare, G. B. Bhanage, J. S. Dhekale, K. G. Dhande, V. V. Aware, *Waste Management*, 2019, **95**, 458-465, doi: 10.1016/j.wasman.2019.06.026.

[4] K. P. Nair, *Tree Crops*. 2021, 1-25, doi: 10.1007/978-3-030-62140-7_1.

[5] H.-S. Kim, K. S. Do, J. H. Park, W. S. Kang, Y. H. Lee, E. W. Park, *The Plant Pathology Journal*, 2020, **36**, 54-66, doi: 10.5423/ppj.oa.11.2019.0281.

[6] M. H. Saleem, J. Potgieter, K. M. Arif, *Precision Agriculture*, 2021, **22**, 2053-2091, doi: 10.1007/s11119-021-09806-x.

[7] G. Fenu, F. M. Malloci, *Big Data and Cognitive Computing*, 2021, **5**, 2, doi: 10.3390/bdcc5010002.

[8] K. Liakos, P. Busato, D. Moshou, S. Pearson, D. Bochtis, *Sensors*, 2018, **18**, 2674, doi: 10.3390/s18082674.

[9] J. Duarte-Carvajalino, D. Alzate, A. Ramirez, J. Santa-Sepulveda, A. Fajardo-Rojas, M. Soto-Suárez, *Remote Sensing*, 2018, **10**, 1513, doi: 10.3390/rs10101513.

[10] S. Chakraborty, R. Ghosh, M. Ghosh, C. D. Fernandes, M. J. Charchar, S. Kelemu, *Plant Pathology*, 2004, **53**, 375-386, doi: 10.1111/j.1365-3059.2004.01044.x.

[11] A. Kumar, R. Agrawal, C. Chattopadhyay, *Mausam*, 2013, **64**, 663-670, doi: 10.54302/mausam.v64i4.749.

[12] L. V. Kanan, M. S. Kumar, *International Journal of Scientific Research in Engineering & Technology*, 2021, **2**, 11–15.
[13] S. Sujatha, R. Bhat, S. Elain Apshara, *International Journal of Innovative Horticulture*, 2018, **7**, 27-37.
[14] D. Mukhopadhyay, A. Chougule, An approach to manage ontology dynamically based on web service composition requests, *Proceedings of the CUBE international information technology conference*, 2012.
[15] A. Danti, *Procedia Technology*, 2012, **4**, 215-219, doi: 10.1016/j.protcy.2012.05.032.
[16] M. Wrzesie´n, W. Treder, K. Klamkowski, W. R. Rudnicki, *Computers and Electronics in Agriculture*, 2019, **161**, 252-259, doi: 10.1016/j.compag.2018.09.026.
[17] A. Sharma, S. K. Mishra, H. Kumar, *Journal of Agrometeorology*, 2017, **19**, 234-238.
[18] L. E. de Oliveira Aparecido, G. de Souza Rolim, J. R. da Silva Cabral de Moraes, C. T. S. Costa, P. S. de Souza, *International Journal of Biometeorology*, 2020, **64**, 671-688, doi: 10.1007/s00484-019-01856-1.
[19] D. Henderson, C. J. Williams, J. S. Miller, *Plant Disease*, 2007, **91**, 951-956, doi: 10.1094/pdis-91-8-0951.
[20] Y. H. Gu, S. J. Yoo, C. J. Park, Y. H. Kim, S. K. Park, J. S. Kim, J. H. Lim, *Computers and Electronics in Agriculture*, 2016, **130**, 169-176, doi: 10.1016/j.compag.2016.10.005.
[21] G. Fenu, F. M. Malloci, *Proceedings of the 2019 3rd International Conference on Big Data Research*, 2019, 76–82, doi: 10.1145/3372454.3372474.
[22] G. Fenu, F. M. Malloci, *International Conference on Intelligent Decision Technologies*, 2020, 79–89, doi: 10.1007/978-981-15-5925-97.
[23] R. Mehmood, P. Sengupta, *9th international conference on computing, communication and networking technologies*, ICCCNT 2018.
[24] C. Boukis, A. Pnevmatikakis, L. Polymenakos, *Artificial intelligence and innovations 2007: from theory to applications*, 2007.
[25] N. El Aboudi, L Benhlima, *2016 International Conference on Engineering MIS (ICEMIS)*, 2016, 1–5, doi: 10.1109/ICEMIS.2016.7745366.
[26] I. Staff, *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*, 2013
[27] P. A. Paul, G. P. Munkvold, *Ecology and Epidemiology*, 2005, **95**, 388-396, doi: 10.1094/phyto-95-0388.
[28] B. Puneeth, P. Nethravathi, *International Journal of Applied Engineering and Management Letters*, 2021, **5**, 183–204, doi: 10.5281/zenodo.5773853.
[29] Arecanut Diseases, 8 Mar 2022, https://vikaspedia.in/agriculture/crop-production/integrated-pest-managment/ipm-for-commercial-crops/ipm-strategies-for-arecanut/arecanut-diseases.
[30] J. Han, A. K. Tung, J. He, *Data Mining for Scientific and Engineering Applications*, 2001, **2**, 461-485, doi: 10.1007/978-1-4615-1733-7_25.
[31] N. David, K. Dimitrios, K. Argyris, S. D. Natasa, P. Pau, C. Roberto, *BMC Bioinformatics*, 2021, **20**, 1-16, doi: 10.1186/s12859-019-3065-1.
[32] C.-H. Wu, J.-M. Ho, D. T. Lee, *IEEE Transactions on Intelligent Transportation Systems*, 2004, **5**, 276-281, doi: 10.1109/tits.2004.837813.
[33] M. Nabipour, P. Nayyeri, H. Jabani, S. Shahab, A. Mosavi, *IEEE Access*, 2020, **8**, 150199-150212, doi: 10.1109/access.2020.3015966.
[34] S. S. Rathore, S. Kumar, *ACM SIGSOFT Software Engineering Notes*, 2016, **41**, 1-6, doi: 10.1145/2853073.2853083.
[35] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, H. Arshad, *Heliyon*, 2018, **4**, e00938, doi: 10.1016/j.heliyon.2018.e00938.
[36] G. Chandrashekar, F. Sahin, *Computers & Electrical Engineering*, 2014, **40**, 16-28, doi: 10.1016/j.compeleceng.2013.11.024.
[37] Narendra, Fukunaga, *IEEE Transactions on Computers*, 1977, **26**, 917-922, doi: 10.1109/tc.1977.1674939.

Author Information

***Ms. Rajashree Krishna**, is Assistant Professor in the Department of Computer Science & Engineering at Manipal Institute of Technology, Manipal, MAHE. Her areas of research interest are Machine learning and Deep learning. She has 15 years of teaching experience, 3 years of research experience and has published papers in Conferences.*

***Dr. Prema K V**, is a Professor and Head in the Department of Computer Science & Engineering at Manipal Institute of Technology, Bangalore, MAHE. Her areas of research interest are Soft computing, Computer Networks, and Security. She has 30 years of teaching experience, 22 years of research experience and has published more than 120 papers in reputed Journals and Conferences.*

***Mr. Rajat Gaonkar** is a final year student in the Department of Electronics and Communication Engineering PES University, Banglore. His areas of research interest are Machine learning and Deep learning.*

