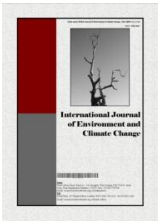

International Journal of Environment and Climate Change

11(2): 32-43, 2021; Article no.IJECC.62599
ISSN: 2581-8627
(Past name: British Journal of Environment & Climate Change, Past ISSN: 2231–4784)

Classification of Level of Rice Pest Incidence Based on Weather Information

V. Jinubala[1*] and P. Jeyakumar[2]

[1]STPI-Software Technology Parks of India, Hyderabad, India.
[2]ICAR-Indian Institute of Rice Research, Hyderabad, India.

Authors' contributions

This work was carried out in collaboration between both authors. Author VJ designed the study, analyzed the data and wrote the first draft. Author PJ assisted in the data collection & analysis, collection of literature and correction of the first draft. Both authors read and approved the final manuscript.

Article Information

DOI: 10.9734/IJECC/2021/v11i230359
Editor(s):
(1) Dr. Wen-Cheng Liu, National United University, Taiwan.
Reviewers:
(1) Ir.Satryo Budi Utomo, University of Jember, Indonesia.
(2) Yu-Chi Pu, National Kaohsiung University of Science and Technology, Taiwan.
(3) Bashir Garba Muktar, Federal University Dutse, Nigeria.
Complete Peer review History: http://www.sdiarticle4.com/review-history/62599

Original Research Article

Received 02 September 2020
Accepted 08 November 2020
Published 17 April 2021

ABSTRACT

Aims: To classify the rice pest data based on the weather attributes using a machine learning approach, a decision tree classifier, and to validate the performance results with other existing techniques through comparison.
Design: Rice pest classification using C5.0 algorithm
Methodology: We collected rice pest data from the crop fields of various regions in the state of Maharashtra of India. The dataset contains the name of the region (Taluk), period (week), pest data, temperature, rainfall, and relative humidity. The data is collected from 39 taluks within four districts in different weeks of the year of 2013-2014. The weather information plays a vital role in this rice pest data analysis, because based on the weather, pest infestation varies in all the regions. The pests considered in this research are Yellow Stem borer, Gall midge, Leaf folder, and Planthopper. The collected dataset is given as input to the classifier, where 75% of data from the dataset is used for training, and 25% of data are used for testing the classifier.
Results: The proposed C5.0 algorithm performed better in the classification of rice pest dataset based on weather attributes. The C5.0 algorithm achieved 88.99% accuracy, 78.81% sensitivity,

**Corresponding author: E-mail: vjinubala@yahoo.com;*

and 89.11% specificity, which are higher in performance when compared with other techniques. Compared with the other different methods, the C5.0 algorithm achieved 1.3 to 8.5% improved accuracy, 2.4 to 9% improved sensitivity, and 0.8 to 7.8% improved specificity.
Conclusion: Early detection of pest and pest based diseases is an essential process to avoid major crop losses. The proposed classification model is designed to classify the level of pest infestations based on weather attributes, as level of infestations caused by the rice pest varies based on weather conditions. The C5.0 algorithm classified the rice pest data based on the weather attributes in the dataset.

Keywords: Rice crop; pest; stem borer; gall midge; leaf folder; planthopper; C5.0 algorithm.

1. INTRODUCTION

Rice is an essential crop of India and its significance in the nation cannot be discredited. India is not just a major rice consumer, but additionally the second-biggest producer after China globally. Pest control management is a major issue for farmers around the world in the agriculture field. Different varieties of rice are cultivated in India like Basmati, Brown, White, Jasmine, Red, Parboiled, and Sticky Rice. Rice is an adaptive crop and can be cultivated in different seasons [1]. The total rice production in India, during 2019-20 is 117.47 million tons. The worldwide losses because of insect pests have reduced from 13.6% in the post-green revolution period to 10.8% towards the start of this century. In India, the crop losses have decreased from 23.3% in the post-green revolution period to 15.7% currently. As far as financial estimation, currently, Indian agriculture endures a yearly loss of about 36 billion US$ [2].

Early detection of crop infections and pests is one of the significant difficulties in the agriculture field. As per the International Rice Research Institute, approximately 37% rice crops losses happen due to pests annually. Out of 266 insect species discovered in rice eco-systems, 42 species are identified as pests. Discovering the diseases or pest and percentage of the disease or pest proportion plays a significant part in the effective cultivation of crops. It is essential to design an agricultural pest classification framework dependent on machine learning-based technology to effectively recognize and target regulatory actions to avoid losses made by pests. The automated method of predicting crop pest incidence reduces huge work of observing large farms, and at the initial stage, disease symptoms are identified [3].

Table 1. List of Pests based on stages

Stages	Pests
Nursery Stage	Stem-borer, Thrips, Gall midge, Root nematode, Root-knot nematode, and Whitetip nematode
Vegetative Stage	Stem-borer, Green leafhopper, Leaf folder, Gall midge, Hispa, Whorl maggot, Mealybug, Case worm
Reproductive Stage	Stem-borer, White-backed planthopper, Brown planthopper, Leaf folder, Green leafhopper, Ear-cutting caterpillar/Cutworm, Gundhi bug, Leaf/Panicle mite

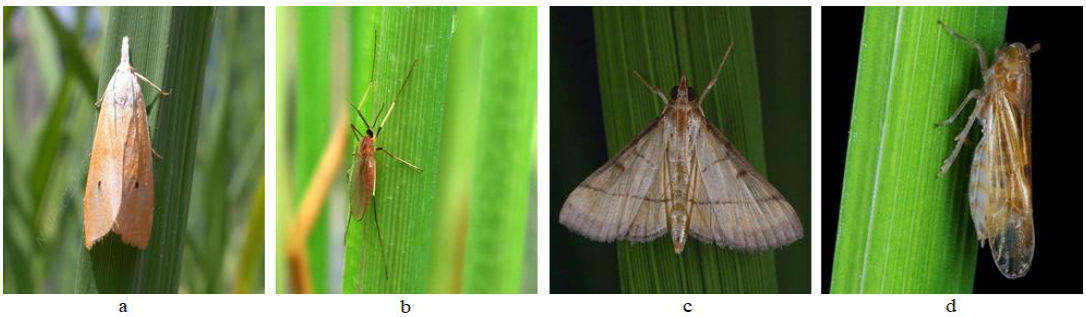

Fig. 1. Rice Pests a) Yellow Stem borer, b) Gall midge, c) Leaf folder, d) Planthopper

1.1 Stem-borers

The stem borers are commonly considered as the most dangerous rice pests globally, they emerge and infect crops from the sowing stage to development [4]. Yellow stem borer (YSB) is the most boundless, prevailing, and dangerous. They are wide spread majorly in the tropics, additionally occur in the climate regions, where the temperature stays over 10 ºC, and more than 1,000 mm rainfall annually.

1.2 Egg Masses

Eggs are laid in masses, usually on leaf tips, and covered with hairs. The quantity of eggs differs in various species.

1.3 Dead Heart

During the vegetative stage, stem borer larva pierces into the stem resulting in the death of central whorl. This is called "Dead heart." The central whorl of affected tillers becomes dry and brownish while the lower leaves stay healthy and green.

1.4 Rice Planthoppers

Numerous species of planthoppers are serious rice pests globally. In several regions, they regularly emerge in numbers sufficiently huge enough to cause total drying of the plant; however, also small populations decrease rice yields. Besides direct feeding damage, planthoppers are vectors of viral diseases such as ragged stunt, grassy stunt, and wilted stunt.

1.5 Gall Midge

Gall midge is a crucial pest of irrigated or rain-fed shallow low land rice. The pest damages majorly during the rice crop tillering stage. The larvae cause damage by feeding on the growing tip lead to the extension of the leaf sheath, which is known as "Gall." The gall similar to an onion leaf glistens in the field, thus it is frequently called a "silver shoot." Profuse tillering and dwarfing of plants are related to the formation of gall.

1.6 Leaf Folders

Leaf folders are broadly spread in the rice-growing regions of 29 humid tropical and temperate nations in Asia, Australia, Africa and Oceania. Damage is observed in all the phases of crop growth. Larva rolls the leaf and feeds by scratching the green matter remaining inside the fold. This feeding results in the growth of longitudinal white streaks. Wherein extreme infection, the leaf tips and edges are dried, and the crop becomes an appearance of whitish burnt. Usually, a single larva is discovered in every fold [5-8].

In this work, the focus of the research is to design a pest classification model based on weather conditions using machine learning technique and the novelty of this research is, the proposed model is developed on using a new decision tree classifier called C5.0 algorithm for classification. The C5.0 algorithm is the advanced technique of C4.5 algorithm, which has lot of advantages over C4.5 algorithm.

1.7 Literature Review

In this present world, each and every domain has obtained its very own form of advancement and development by the research made by scientists and scholars worldwide. Agriculture is one of the main fields to concentrate and it requires many innovations to solve field oriented issues to support farmers. Pest control is one of the primary issues faced by farmers all over the world and many researchers and authors has provided solutions based on artificial intelligence techniques like machine learning and deep learning. For pest control and classification model, the related works are analyzed and utilized in the proposed model.

Jinubala et al. [9] proposed a model for Rice Pest data classification utilizing the Decision Tree algorithm. The decision tree classification methods and the primary issues in classification and predicting techniques for agricultural data were analyzed. Different classification methods have been applied with the Leaf Folder pest dataset of rice crop to classify them into four-classes dependent on pest intensity range during the entire crop season, utilizing R statistical language. The classification algorithms were tested with the Iris Flower benchmark dataset, and the performance of decision tree classification algorithms was tested before applying the rice pest dataset. At last, the classification algorithms were tested with rice crop pest dataset, and the performance was evaluated utilizing classification accuracy. Out of six classification techniques, it was discovered that C4.5 (decision tree) was efficient with a classification accuracy of 78% [9].

Ahmad Arib Alfarisy et al. [10]; analyzed a new technique utilizing deep learning for classifying paddy pests and infections. Among many advanced deep learning systems, the Caffe model was used. It is an open-source deep-learning system created by BVLC, and a model of pre-trained CaffeNet, combined with Caffe was additionally utilized in this analysis to check the practicality of using automated tools for supporting paddy farmers for identifying diseases and pests those are damaging their paddy field. The proposed model classified 13 types of paddy pests and infections with an accuracy of 87% [10].

Takuya Kodama and Yutaka Hata [11]; developed a system for classifying healthy rice crops and diseased rice crops by processing images from rice seeded in the paddy field. Since the symptom indeed occurs in rice diseases, the color data was considered. In this way, the pixel value was utilized as the feature, and a classifier using SVM was analyzed. Also, the learning time was decreased by using the principal component analysis. Hence, the model acquired over 90% accuracy [11].

Chowdhury Rafeed Rahman et al. [12]; presented deep learning-based methods for detecting pests and infections in rice crops utilizing images obtained in real-time conditions. Different architectures of CNN experimented on the big dataset of rice pests and diseases gathered manually from the field, which includes both intra and inter-class varieties and has 9 classes altogether. The outcomes demonstrated that rice pests and diseases could be detected efficiently utilizing CNN with 99.53% accuracy on the test set using the CNN model, VGG16. These models were not appropriate for mobile because of their vast size having more parameters. So, a new architecture of CNN was proposed called stacked CNN that used two-stage training to minimize the model's size substantially while simultaneously keeping up high classification accuracy. The test outcomes represented that test accuracy of 95% was obtained with stacked CNN while decreasing the size of the model by 98% contrasted with VGG16. This sort of memory-efficient CNN models could support rice pests and mobile application development based diseases detection [12].

Eusebio L. M Jr. and Thelma D. P, [3]; developed an application that could support farmers in identifying pests and infections of rice utilizing CNN and image processing. By implementing CNN and image processing, the app was developed to detect rice pest and infections. The searching and correlation of collected images to a stack of rice pest images were experimented utilizing a CNN model. Compiled images were preprocessed and were used in training. Cross-entropy was less, which implied that the trained method could execute prediction or classified images with a less error percentage. The model accomplished 90.9% test accuracy finally.

Manikandan N et al. [13]; proposed a model of weather-based pest prediction for major rice insect pests. Weather-based pest prediction systems were generally used in the coordinated pest management framework as a tool that does not harm the predators and reduces environmental pollution. On this basis, an effort was made for predicting the population prevalence of YSB, BPH, and RLF (Rice Leaf organizer). Generalized Linear Model (GLiM) was implemented for anticipating the population of YSB, BPH, and RLF. The outcomes of the chi-square test uncovered that numerous different factors that impact the number of light trap catches of the insects separated from climate parameters. The equation predictability could be expanded if the climate factors were integrated with different elements (soil, variety, fertilizer application, and so on.) in the model development [13].

2. MATERIALS AND METHODS

Weather-based classification models are broadly used as a tool in the integrated pest management system. Based on this, a model was designed to classify the prevalence of Yellow Stem borer, Gall midge, Leaf folder, and Planthoppers on different weather conditions. The machine learning model considered for this analysis is the C5.0 algorithm, which is a Decision Tree (DT) based classifier algorithm. Contrasted with more developed and advanced machine learning models, the decision trees under the C5.0 algorithm usually perform better in several cases and simple to understand and implement.

2.1 C5.0 Algorithm

The DT is a technique to perform the classification. DTs have turned out as the most effective and notable techniques in machine learning and data mining. DTs need two sorts of information: training and testing. Training, which is commonly the huge data, is used to create mining trees. The testing data is used to get the accuracy and misclassification rate of the DT.

The flowchart of the proposed model is shown in Fig. 2, in which the collected dataset from various regions is given as input to the model. The dataset contains different parameters like pest information, location of the data collected, maximum and minimum temperature of that location, rainfall of that location and humidity of that location. These attributes are selected as features to train the proposed classifier model and the dataset is divided into 75% for training and 25% for testing. After the training, the performance analysis is carried out by testing the dataset and the classified results will be based on the infection caused by the pests from different regions. The operation of the C5.0 algorithm is discussed in the following paragraphs.

C5.0 algorithm is a development of the C4.5 algorithm, which is an expansion of the ID3 algorithm. It is a classification algorithm that is suitable for large datasets. It is better than C4.5 on efficiency, memory, and speed [14]. C5.0 algorithm uses a pruning technique. Once a DT is built, some branches may contemplate errors in the training data because of noise that is expelled by the tree pruning techniques. The tree pruning technique uses the statistical measure to expel the least reliable branches. Post-pruning and Pre-pruning are the two usual methods. In the pre-pruning technique, the tree is pruned by determining not to additionally split the sub set of training tuples at a presented node. Post-pruning process expels subtrees from a fully grown tree, by exchanging a subtree with a leaf labelled as the most prevalent class in it. C5.0 also uses a BOOSTING method to generate and integrate various classifiers to deliver enhanced predictive accuracy [15]. Contrasted with C4.5, the error rate of the C5.0 classifier is around 1:3 of the C4.5 classifier. Ross Quinlan developed the C5.0 algorithm with solutions for the classification issues. C5.0 works in three essential stages; initially, all samples are considered at the top of the tree called as the root node and forwarded them through to the second node called "branch node." The branch node produces rules for a set of samples reliant on an entropy measure. At this point, the C5.0 creates a huge tree by considering all characteristic values and concludes the decision principle by pruning. It uses a heuristic technique for pruning depended on the statistical measure of splits. Hence fixing the better rule, the branch nodes send the last

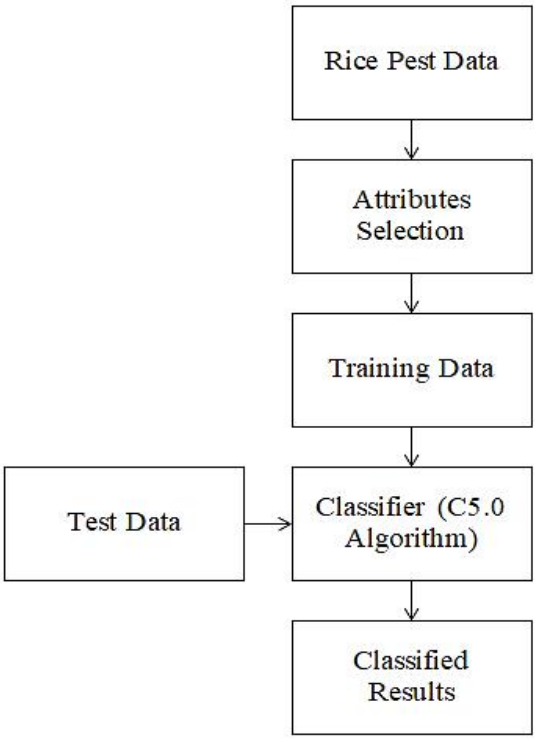

Fig. 2. Proposed model

class value in the previous node, known as the "leaf node." Machine learning enables to predict an outcome using data about past events. C5.0 algorithm is used to develop a DT for classification.

There are two sequences, $\mathcal{X} = \{X^1, X^2, \dots, X^N\}$ is a training dataset and $Y = \{y_1, y_2, \dots, y_n\}$ is a set of corresponding classes. Here $X^i = \{x_1^i, x_2^i, \dots, x_d^i\}$ is a vector of attributes, where $i \in \{1, \dots, N\}$, d is total attributes, N is total vectors in the training dataset, $y_i \in C = \{1, \dots, M\}$ is a number of class of X^i vector. An attribute x_j^i is a discrete or a real-valued variable, i.e. $x_j^i \in \{1, \dots, T_j\}$ for some integer T_j. Let $DOM_j = R$ if x_j is a real value; and $DOM_j = \{1, \dots, T_j\}$ if x_j is a discrete-valued attribute. The problem is to develop a function $F: DOM_1 \times \dots \times DOM_d \to C$ that is a classifier. The function classifies a new vector $X = (x_1, \dots, x_d)$ that is not from $\mathcal{X}$.

DT is a tree so that every node tests some condition on input variables. Assume B is some test with outputs $b_1, b_2, \dots, b_t$ that is tested in a node. Then, there are *t* outgoing edges for the node for every output. Every leaf is linked with a result class from *C*. The process of testing is as follows; test conditions are initiated from the root node and pass by edges based on a result of the condition. The label on the reached leaf is the output of the process of classification. In a training set, the number of vectors is *N*; the number of classes is *M*, and the number of attributes is *d*. Let the height of a building tree *h* be the algorithm's parameter. Let *RVA* be the real-valued attribute's set indexes, and *DVA* is the discrete-valued attribute's set indexes.

The following will be the procedure for enhancements. Assume that a binary tree was constructed of height *h*. The primary process is Construct-Classifiers that implement a recursive process, Form-Tree for developing nodes. The essential parameters of Form-Tree are level that is an index of tree-level; the tree that is a result subtree that the procedure will construct; X' that is a set that used for constructing this subtree. The Form-Tree process performs two steps. The first one Choose-Split is selecting the test *B* that is the selecting an attribute and the splitting by this attribute that expand the objective function $\frac{G(X';B)}{P(X';B)}$. The result attribute index is *att,* and the resulting split is a split variable. The second step Divide is the splitting process itself.

Algorithm: Construct-Classifiers and Form-Tree
procedure Construct-Classifier( )

$X' \leftarrow X, level \leftarrow 1$

FormTree (tree, level, X')
end procedure
procedure Form-Tree (tree, level, X')
att, split ← ChooseSplit (X')
Divide (tree, attr, split, level)
end procedure

2.2 Proposed C5.0 Algorithm

Input: a random-valued dataset *DS*
Output: Optimal *Tree*
Tree = *Φ*;
if *DS is "pure"* **then**
STOP;
end
for *every attribute 'a' belongs to DS* **do**
Correlate data-theoretic standards if we split on 'a';
end
a_{best} = best attribute based on the above-assessed standard;
Tree = generate a decision node that tests a_{best} in the root;
D_v = induced sub-datasets from DS based on a_{best};
for *all* D_v **do**
$Tree_v$ = C5.0(D_v);
Attach $Tree_v$ to the equivalent branch of Tree;
end
return *Tree*;

- *A* - Attribute.
- *DS* - Total dataset.
- *T* - An attribute with *n* number of mutually exclusive outputs T_1, T_2 ... T_n.
- *c* - Count of classes.
- *p*(*DS, j*) - the ratio of instances in DS according to the j^{th} class.
- $D_i \subseteq DS$ – the division of dataset where each record has T_i for the attribute *T*.
- $|D_i|$ - size of the division D_i.

C5.0 algorithm initially computes the entropy of the total dataset (*DS*) as follows,

$$I(DS) = -\sum_{j=1}^{c} p(DS, j) log_2\big(p(DS, j)\big) \qquad \text{(a)}$$

If *T* is a categorical variable, C5.0 algorithm in the following stage computes entropy inside a dataset where each record has T_i for *T*. It calculates the entropy as,

$$I(D_i) = -\sum_{j=1}^{c} p(D_i, j) log_2\big(p(D_i, j)\big) \quad \text{(b)}$$

Hence, the total dataset's weighted entropy when variable *T* is analyzed at the initial node is as,

$$I(DS, T) = -\sum_{i=1}^{n} \frac{|D_i|}{|DS|} \times I(D_i) \quad \text{(c)}$$

If a numerical attribute *T* holding a domain [*l, u*], hence the data were initially re-organized; thus, *T* values were arranged in descending or ascending order. Then the dataset was separated with two divisions D_1 and D_2 depend on a split point *p*; thus, the *T* in D_1 was [*l, p*], and D_2 was [*p* +1*, u*], where *p* + 1 refers the following high value to *p* in the domain. Then *I* (*DS, T*) is computed as,

$$I(D_i) = -\sum_{j=1}^{c} p(D_i, j) log_2(p(D_i, j)), for 1 \leq i \leq 2 \quad \text{(d)}$$

$$I(DS, T) = -\sum_{i=1}^{2} \frac{|D_i|}{|DS|} \times I(D_i) \quad \text{(e)}$$

I (*DS, T*) are computed for every feasible split points of *T*. At last, the minimal *I* (*DS, T*) was deemed as the *I*(*DS, T*) of *T* and the split point which generates the minimal *I* (*DS, T*) was considered to be as the best *T* split point.

The entropy reduction, through selecting an attribute *T* as a test variable, was deemed as the data gain for the variable and computed as,

$$Gain(DS, T) = I(DS) - I(DS, T) \quad \text{(f)}$$

Gain (*DS, T*) of *T* was impacted due to the size of the domain of *T* and would be maximal while there was just a single record in every subset D_i. Thus, the discussed gain computation supports the attribute with a large domain size over those holding small size. To decrease this excessive favor, the gain ratio of the attribute was used to choose the test variable for the node. The gain ratio was computed by,

$$Gain\ Ratio(DS, T) = \frac{I(DS) - I(DS, T)}{Split(DS, T)} \quad \text{(g)}$$

The split data of attribute *Split* (*DS, T*) expands once the attribute has a larger size domain. Split data of every attribute was computed as,

$$Split(DS, T) = -\sum_{i=1}^{k} \frac{|D_i|}{|DS|} \times log_2 \frac{|D_i|}{|DS|} \quad \text{(h)}$$

Where, the domain size of *T*, |*T*| = *k*.

At last, the one holding the maximum gain ratio was selected as the root node of DT from total non-class attributes. If the selected variable *T* is a categorical variable containing the size of domain |*T*|= *k*, hence the dataset *DS* was isolated into *k* conflicting partitions D_1, D_2 ... ... D_k. Instead, if the selected variable *T* was a numerical variable with domain [*l, u*] therefore the dataset *DS* was isolated into D_1 and D_2 divisions using the best split point of *T*. When the test variable of a DT is selected for the root node, the similar operations are reiterated frequently on every division of the dataset till an end condition.

The proposed classification model is designed to classify the pest based on weather conditions by using a machine learning-based classifier called the C5.0 algorithm. This classification model works in simple steps, as the initial step is to analyze the dataset, and the attributes are selected like temperature, rainfall, and period. Then the classifier needs to train on the dataset with 75% of the data, and for testing, 25% of the data is used. In the final step, based on the training and testing the classifier can detect the pest information based on weather conditions. The result is based on which pest is causing more infection on which weather conditions.

3. RESULTS AND DISCUSSION

As per the Indian Meteorological Department, in India there are four climatological seasons:

Winter: December to February. The average temperature is around 10-15°C in northwest and 20-25°C in southeast.
Summer or Pre-monsoon: March to May. Average temperature is around 32-40°C in most of the regions.
Monsoon or Rainy: June to September.
Post-monsoon or autumn: October to November.

3.1 Dataset Description

The rice pest dataset is collected from various areas (Talukas) of State of Maharashtra of India. The dataset contains the name of the area (Taluk), period (week), pest information, temperature, rainfall, and relative humidity. The data is collected from 39 talukas within four districts in different weeks of the year of 2013-2014. The weather information plays a vital role in this dataset. Based on the weather, pest information varies in all the regions. The pests considered in this research are Yellow stem borer, Gall midge, Leaf folder, and Planthopper. The dataset sample from every 39 taluks is shown in Table.2.

Table 2. Sample dataset

Taluk	Taluk wise rice pest 2013-14 without zero rows week	Stem borer dead heart (No. per hill)	Stem borer egg mass (No. per hill)	Gall midge damage (No. per hill)	Leaf folder folded leaf (No. per hill)	Plant hoppers (No. per hill)	Min temp	Max temp	Rain fall	Relative humidity
Bhandara	35	0.154688	0	0.000781	0.213281	0.092188	24	34	18.4	82
Lakhandur	30	0.078226	0.020161	0.021774	0.08629	0	24	27	120.3	79
Lakhani	36	0.106563	0.016563	0.051719	0.175625	0.020781	25	33	12	79
Mohadi	34	0.089474	0.014737	0.000526	0.204737	0.001053	24	0	65.4	81
Pauni	44	0.232258	0.074194	0.002419	0.175	1.497581	20	0	0	61
Sakoli	37	0.164063	0.019531	0.050781	0.094531	0.208594	25	34	14.7	78
Tumsar	38	0.407031	0.003906	0.2625	0.303906	1.111719	24	32	66.9	75
Ballarpur	41	0	0	0	0.005882	0	24	30	44.5	95
Bhadravati	40	0.183333	0	0.141667	0.2	0.275	24	30	122	92
Bramhapuri	42	0.05	0.01	0	0.16	0.32	23	32	0.1	83
Chimur	43	0.08	0	0.006154	0.071538	0.267692	24	30	30.4	89
Gondpimpri	43	0.038889	0	0	0.136111	0.044444	24	30	30.4	89
Korpana	46	0	0	0	0.05	0.05	15	26	0	71
Mul	40	0	0	0	0.041176	0.0125	24	30	122	92
Nagbhid	45	0.012121	0	0.013636	0.024242	0.125758	19	28	0	72
Pomphurna	43	0.007143	0	0	0.164286	0.119048	24	30	30.4	89
Rajura	45	0	0	0	0.06875	0.04375	19	28	0	72
Savali	40	0	0	0	0.015625	0.040625	24	30	122	92
Sindevahi	35	0.029688	0.01875	0.054688	0.067188	0.004688	25	30	10.2	89
Aheri	39	0.021875	0	0	0.060938	0.0375	0	0	7.3	69
Armori	37	0.010938	0	0.003125	0.023438	0	0	0	36.3	0
Bhamaragad	39	0.003125	0	0.001563	0.029688	0.03125	0	0	7.3	69
Chamoshri	40	0.011719	0	0	0.052344	0.0375	0	0	101.6	0
Dhanora	44	0.003125	0	0	0.009375	0.003125	0	0	0	0
Etapalli	38	0.010938	0	0	0.014063	0.020313	0	0	117.1	69
Gadchirol	40	0.012308	0	0	0.038462	0.028462	0	0	101.6	0
Korachi	34	0.00625	0	0.039063	0.003125	0	0	0	52.4	81
Kurkheda	31	0.003125	0	0	0	0	0	0	284.2	82
Mulchera	36	0.0125	0	0	0.014583	0.004167	0	0	27.1	78

Taluk	Taluk wise rice pest 2013-14 without zero rows week	Stem borer dead heart (No. per hill)	Stem borer egg mass (No. per hill)	Gall midge damage (No. per hill)	Leaf folder folded leaf (No. per hill)	Plant hoppers (No. per hill)	Min temp	Max temp	Rain fall	Relative humidity
Sironcha	42	0.001515	0	0	0.039394	0.025758	17	0	0.2	0
Vadsa	33	0.021875	0	0.035938	0.01875	0	0	0	207.5	83
Amgaon	29	0.035714	0	0.007143	0.035714	0	0	0	132.8	0
Devrai	36	0.023438	0.002344	0.064844	0.011719	0.004688	23	33	9.8	79
Gondia	40	0.335165	0.01978	0.041758	0.046703	0.195055	22	29	63.7	91
Goregaon	34	0.032813	0.010156	0.023438	0.030469	0.004688	23	26	81.7	81
Morgaoanrjuni	32	0.034615	0	0.007692	0.007692	0	24	29	87.9	83
Sadakarjuni	33	0.325	0.000781	0.032813	0.011719	0	29	29	195.5	82
Salkosa	36	0.070313	0.003125	0.032813	0	0.060938	23	33	9.8	79
Tiroda	43	0.132258	0.026613	0.002419	0.018548	2.047581	20	29	7.7	85

3.2 Performance Metrics

The performance metrics used in this work are accuracy, sensitivity, and specificity. The confusion matrix concept is presented in Table.3.

Table 3. Confusion matrix

	Actual positive	Actual negative
Predicted Positive	TP	FP
Predicted Negative	FN	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \% \quad \text{(i)}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad \text{(j)}$$

$$\text{Specificity} = \frac{TN}{TN+FP} \quad \text{(k)}$$

TP: The count of correct classifications in pest affected class.
FP: The count of incorrect classifications in pest affected class.
TN: The count of correct classifications in the unaffected pest class.
FN: The number of incorrect classifications in unaffected pest class.

From the dataset classification, the stem borer-dead heart with the highest infestation rate is at Chamoshri taluk in the 29th week of the year 2013-2014, stem borer-egg mass with the highest infestation rate is in Pauni taluk in the 42nd week, gall midge with the highest infestation rate is in Tumsar taluk in the 37th week, leaf folder with the highest infestation rate is in Korachi taluk in the 40th week, and planthopper with the highest infestation rate is in Lakhandur taluk in the 42nd week of the year. From the dataset, the minimum temperature is recorded as 12ºC and maximum temperature as 35ºC in the talukas, and the highest rainfall is recorded as 284.2mm in Chamoshri taluk and highest relative humidity up to 95% in few talukas.

The proposed model is designed to classify the pest level based on weather attributes by using the C5.0 algorithm. In this model, initially the dataset was processed, and the attributes like temperature, rainfall, relative humidity and period are selected based on the classifier. Then the classifier is trained with 75% of the dataset, and 25% of the dataset is used for testing. Finally, based on the training and test dataset, the classifier detected the level of pest infestation based on weather attributes. The result is based on the infestation caused by the pests on different weather conditions. The performance analysis is evaluated based on confusion matrix in terms of accuracy, sensitivity, and specificity. The obtained results are compared with different existing machine learning techniques for validation as shown in Table.4.

Decreasing the false positive and false negative in the process is essential in any research. By minimizing false positive and false negative values, it will be very helpful not to misclassify any data that is very relevant in the classification model. A false positive is an error in classification, in which a test result incorrectly specifies the presence of a condition such as a pest infection when the infection is not present, while a false negative is an opposite error where the test result incorrectly fails to specify the presence of a condition when it is present. These kinds of errors lead to reduce the performance of the model in achieving less accuracy, sensitivity and specificity.

Table 4. Performance analysis and comparison of techniques

Sl. No.	Algorithm	Accuracy	Sensitivity	Specificity
1	Naive Bayes	87.62	76.32	88.23
2	C4.5	86.36	75.13	86.44
3	Genetic Algorithm	86.27	75.12	86.14
4	Support Vector Machine	85.25	74.33	85.65
5	Artificial Neural Network	84.31	73.56	85.42
6	Principal Component Analysis	83.34	72.74	84.34
7	Linear Discriminant Analysis	82.55	71.36	83.27
8	Particle Swarm Optimization	81.42	70.99	82.14
9	Dimensionality Reduction	80.43	69.72	81.26
10	C5.0 (Proposed)	88.99	78.81	89.11

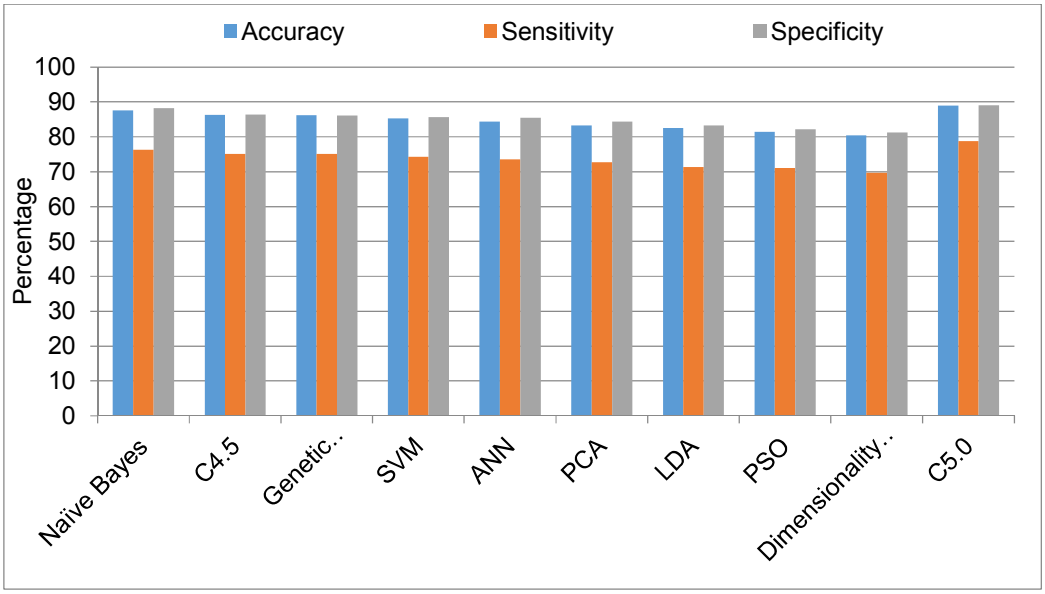

Fig. 3. Graphical representation of performance analysis

The proposed model performed better in the classification of rice pest dataset based on weather correlation. As shown in table.4, the C5.0 algorithm achieved 88.99% accuracy, 78.81% sensitivity, and 89.11% specificity, which are higher in comparison to the performance of the other techniques. Compared with the different techniques, the C5.0 algorithm achieved 1.3 to 8.5% improved accuracy, 2.4 to 9% improved sensitivity, and 0.8 to 7.8% improved specificity.

4. CONCLUSION

The proposed classification model is designed to classify the level of pest incidence based on weather attributes by using a machine learning-based classifier. The machine learning classifier is used to classify the pest dataset collected based on the weather information provided. In this research, the C5.0 algorithm, a decision tree based classifier is used for classification. The dataset is collected from 39 talukas within four districts of Maharashtra State in India, during different weeks of the year of 2013-2014. The pests considered in this research are Yellow Stem borer, Gall midge, Leaf folder, and Planthoppers, which are the major crop damaging pests in the rice crop fields around India. 75% of data from the dataset is used for training, and 25% of dataset is used as test dataset for the classifier. The result is based on the classification parameters like accuracy, sensitivity, and specificity. The C5.0 algorithm achieved 88.99% accuracy, 78.81% sensitivity, and 89.11% specificity, which are higher in comparison to the performance of other techniques. In the future, a hybrid technique can be combined with the proposed model to improve the classification performance by reducing the false positives and flse negatives, which could lead to achieve higher accuracy and improved performance.

ACKNOWLEDGEMENT

The authors are grateful to Commissionerate of Agriculture, Government of Maharashtra, for sharing the pest dataset from crop pest surveillance and advisory project (CROPSAP).

COMPETING INTERESTS

Authors have declared that no competing interests exist.

REFERENCES

1. Noor Ul Ane, Mubashar Hussain. Diversity of insect pests in major rice-growing areas of the world. Journal of Entomology and Zoology Studies. 2002;4(1):36-41.
2. Dhaliwal GS, Jindal Vikas, Mohindru Bharathi. Crop Losses due to Insect Pests: Global and Indian Scenario. Indian Journal of Entomology. 2015;77(2):165-168.
3. Eusebio L. Mique, Jr., Thelma D. Palaoag. Rice pest and disease detection using

Convolutional neural network, ICISS '18: Proceedings of the 2018 International Conference on Information Science and System. ACM Digital Library. 2018;147-151.
4. Pathak MD, Khan ZR. Insect pests of rice, International Rice Research Institute. International Centre of Insect Physiology and Ecology. 1994;1-89.
5. Yang Lu, Shujuan Yi, Nianyin Zeng, Yurong Liu, Yong Zhang. Identification of rice diseases using deep convolutional neural networks. Neurocomputing. Elsevier. 2017;267:378-384.
6. Rajmohan R, Pajany M, Rajesh R, Raghu Raman D, Prabu U. Smart paddy crop disease identification and management using deep Convolution Neural Network and SVM Classifier. International Journal of Pure and Applied Mathematics, 2018;118(15):255-264.
7. Ebrahimi MA, Khoshtaghaza MH, Minaei S, Jamshidi B. Vision-based pest detection based on SVM classification method. Computers and Electronics in Agriculture. Elsevier. 2017;137:52-58.
8. Vennila S, Jitendra Singh, Priyanka Wahi, Manisha Bagri, DK Das, M Srinivasa Rao. Web-enabled weather based prediction for insect pests of rice. ICAR-National Research Centre for Integrated Pest Management. 2016;1-50.
9. Jinubala V, Lawrance R, Jeyakumar P. Classification of rice pest data using decision tree algorithm. International Journal of Research in Advent Technology. 2019;7(5S):148-154.
10. Ahmad Arib Alfarisy, Quan Chen, Minyi Guo. Deep learning-based classification for paddy pests & Diseases Recognition. ICMAI '18: Proceedings of 2018 International Conference on Mathematics and Artificial Intelligence. ACM Digital Library. 2018;21-25.
11. Takuya Kodama, Yutaka Hata. Development of classification system of rice disease using artificial intelligence. In IEEE International Conference on Systems, Man, and Cybernetics. 2018;3699-3702.
12. Chowdhury Rafeed Rahman et al. Identification and Recognition of Rice Diseases and Pests Using Convolutional Neural Networks. Computer Vision and Pattern Recognition. arXiv:1812.01043 [cs.CV], 2019;1-25.
13. Manikandan Narayanasamy, J. S. Kennedy, V. Geethalakshmi. Weather based pest forewarning model for major insect pests of rice – An effective way for insect pest prediction. Annual Research and Review in Biology. 2017;21(4):1-13.
14. Kamil Khadiev, Ilnaz Mannapov, Liliya Safina. The quantum version of classification decision Tree Constructing Algorithm C5.0. ArXiV 2019. ArXiv, abs/1907.06840.
15. Vahid Rafe, Sara Hashemi Farhoud, Siamak Rasoolzadeh. Breast cancer prediction by using C5.0 Algorithm and BOOSTING method. Journal of Medical Imaging and Health Informatics. 2014;4(4):600-604.

Peer-review history:
The peer review history for this paper can be accessed here:
http://www.sdiarticle4.com/review-history/62599