International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
***Article Received**: 15 July 2020 **Revised**: 20 July 2020 **Accepted**: 25 July 2020 **Publication**: 31 July 2020*

A Comparative Study of Ensemble Classifiers for Paddy Blast Disease Prediction Model

Varsha M.
Department of Information Science and Engineering
Bapuji Institute of Engineering and Technology
Davangere, India
varshabiet@gmail.com

Dr. Poornima B.
Department of Information Science and Engineering
Bapuji Institute of Engineering and Technology
Davangere, India
poornimateju@gmail.com

Abstract—Paddy blast has become most epidemic disease in many rice growing countries. Various statistical methods have been used for the prediction of paddy blast but previously used methods failed in predicting diseases with good accuracy. However the need to develop new model that considers both weather factors and non weather data called blast disease data that influences paddy disease to grow. Given this point we developed ensemble classifier based paddy disease prediction model taking weather data from January 2013 to December 2019 from Agricultural and Horticulture Research Station Kathalgere Davangere District. For the predictive model we collected 7 kinds of weather data and 7 kinds of disease related data that includes Minimum Temperature, Maximum Temperature, Temperautre Difference,Relative Humidity, Stages of Paddy Cultivation, Varities of seeds, Season of cropping and so on. It is observed and analyzed that Minimum Temperature, Humidity and Rainfall has huge correlation with occurrence of disease. Since some of the variables are non numeric to convert them to numeric data one hot encoding approach is followed and to improve efficiency of ensemble classifiers 4 different filter based features selection methods are used such as Pearson's correlation, Mutual information, ANNOVA F Value, Chi Square. Three different ensemble classifiers are used as predictive models and classifiers are compared it is observed that Bagging ensemble technique has achieved accuracy of 98% compared to Adaboost of 97% and Voting classifier of 88%. Other classification metrics are used evaluate different classifiers like precision, recall, F1 Score, ROC and precision recall score. Our proposed ensemble classifers for paddy blast disease prediction has achieved high precision and high recall but when the solutions of model are closely looked bagging classifier is better compared to other ensemble classifers that are proposed in predicting paddy blast disease.

Keywords- Paddy Blast Disease, Mutual Information, ANNOVA F Value, Voting Classifer, Bagging, Adaboost, Precision-recall Score, ROC

I. INTRODUCTION

Rice is the main food crop of India and one of the country's utmost critical agricultural commodities. It is estimated that Paddy has occupied the principal position in Indian Agriculture with 24% of gross cropped area of the country. It contributes total of 42% of total food grain production and 45% of total ceralac production of the country. Karnataka is the one amoung major rice growing states in the India. In Karnataka state it is grown under a variety of soil, temperature and Rainfall .

Production of paddy has been challenged task nowdays by the recent changes in production of crop and occurrence of disease. Extensive management of crop including large use fertilization, flooding and further unintrupted mixed farming helps paddy in development of pathogens from one crop toanother. Host Presence, Infected Seeds, and unclean cultivation are also major factors of disease spreading. Close planting of paddy along with increase in humidity and decrease in temperature acts as a favourable condition for disease to spread. Loss of paddy production are also due to diseases diveresed depending on season, environmental conditions and various cultivation practices.

Paddy yields in India fluctuate from year to year repeatedly as a result of uneven climate or sometimes due to destruction from major diseases and pests. Atmospheric weakness resulting from a changing climate is greatly correlated with the development of Paddy diseases. Owing to the degree and range of these changes are very indefinite, the forecasting of

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
***Article Received**: 15 July 2020 **Revised**: 20 July 2020 **Accepted**: 25 July 2020 **Publication**: 31 July 2020*

Atmospheric change effecting on these rice pathosystems is difficult and unproven.

Paddy blast, precipitate by the fungus M. grisea is the major of all Paddy diseases and is scattered throughout the globe. It can damage any organ of rice plant and the plants get the maximum disease at tillering stage . Blast is found in average of 80 countries throughtout the globe. Its first incident was in china and was known as fever disease .

Paddy blast causes very huge loss of upto 90% and above. Losses of 20-25% intensity were recorded in Mandya District of Karnataka(2013-2015), 26-50% intensity were recorded in Mysore District(2013-2015), 5-10% intensity in Shivamogga,Davangere and Bellary Districts (2013-2015) [1].

For the Previous 20 years magnificient forwards have been made in studying Blast Disease. Problem of disease is mainly in Kharif Season even disease has equal proportion in irrigated region also . This blast disease is also associated with nutrition imbalances when nitogen nutrient is less.

Forewarning model based on the relationship between the atmospheric conditions , variety of seeds grown in particular region report date of occurrence of same disease in previous years severity of the disease in previous years could be used to control management results. If a good prediction model is developed thus occurrence of disease could be avoided by suitable application of the control measures. Developing such predictive models will help to reduce cost of paddy production by reducing extensive use of chemical fertilizers. Since environmental parameters plays vital role in the appearance, multiplication and spread of the blast disease a forewarning system provides the desired results to consumers.

In this research work we developed predictive model for investigating environmental factors, varities of seeds, growth stages, report date, severity that corelates with the occurrence of blast disease in the region of davangere district Karnataka state. Ensemble classifiers were built to forecast whether or not blast disease will occur for the given atmospheric factors and other conditions. For this study we have used 7 year Weather data and disease related data collected from University of Agriculture and Horticulture Science Kathalgere Davangere Taluk Davangere District This proposed system helps in early prediction of the occurrence of paddy blast and contributes to the disease management. The proposed model is developed for Davangere district of Karnataka state considering all factors that influence blast disease to grow.

The further content of this is as follows: The related work is discussed in section 2. In section 3 Materials and Methods are discussed where data sets were collected and used to build model. In section 4 Feature Selection Methods and Ensemble Classifiers are described followed by this section 5 discusses results finally section 6 concludes the paper.

II. RELATED WORK

Shafaullah, Muhammad[1] conducted experiment at University of Agriculture Faisalabad and explored the effect of epidemiological factors such as temperature, humidity and rainfall during the year 2008. Study found that tempearute and blast disease was negatively corelated i.e -0.88, humidity was positively corelated with paddy blast with 0.95 of correlation and rainfall also positively correlated thus describes as paddy blast disease incidence increases with the decrease in temperature, increase in humidity and increase in rainfall.

Rakesh Kaundal[8] selected six weather variables as predictor variables two series of model were developed called as cross year and cross location models. And developed model also validated using five fold cross validation procedure separately with regression , back propogation neural network, generalized regression neural network and SVM and in the study it is found that SVM based prediction model performed well compared to other models in the study.

Yoshihiro Taguchi[3] studied effect of fan-forced wind on paddy blast disease in two seasons. Author considered a Paddy field (45m * 15m) with history of low incidence of paddy disease at refereed this field as G2 fileld. And Fields adjacent to G2 towards north direction considered as control field and it is referred as C1 and C2 Fields by doing this authors have observed that blast fungus will be active when the humidity is high that is 90% and above and it reaches to peak in the midnight using fanforced wind wetness period can be decreased thus this prevents blast pathogens penetration level.

Kwan-Hyung Kim[4] evaluated EPIRICE model with historical disease incidence data and weather data from 2002-2010. This evaluation is carried out in South Korea. EPRICE model calibrated and validated against observed disease incidence data for leaf blast and sheath blight. The results are compared using disease progress curve and area under disease progress curve. Accepatance Test were applied to the model to check accuracy of EPIRCE. Further IPCC and RCP scenarios were used as inputs to the model output of the model displayed using GIS to show possible changes for both the diseases.

Rupankar Bhagawati[6] aimed to review the scenario of climatic change and its impact on productivity of agriculture and farming system. Region witnessed rise in minimum temperature, maximum and minimum rainfall remains constant there can be chances of incidence of disease on crop.

Rajendra Prasad[7] presented work that focused on study of weather factors on rice bast in mid hill conditions of Himachal Pradesh region. Region specific field experiments conducted through 1984 to 2012 at Palampur. Two varities of seeds were cropped namely Hasan,sarai,china 988 leaf symptoms were observed during mid of tillering stage when coincided with weather parameters for all the years and authors also observed that minimum temperature 20 degree C rainfall and more cloudy days influence incidence of blast disease.

Y.H. Gu , S.J. Yoo[9] developed the potato late blight prediction model called as BLIGHT-SVR. After model predicted and it is also verified the first date of disease occurrence during the year 1976 to 1985 and through 2009 to 2010 Regression analysis was conducted called Support vector Regression(SVR) that offered better performance 13 different kinds of wathers factors were analysed including temperature, humidity, evaporation and so on these factors are highly correlated with first date of disease occurrence. Accuracy of prediction model was 64.3% and it showed highest accuracy compared to pace regression and linear regression.

Jia-You Hsieh[10] developed prediction model considering five year environmental data from 2014 to 2018 these data used as candidate data collected by the taiwan government. Labels are assigned via field observations. Authors applied

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
***Article Received**: 15 July 2020 **Revised**: 20 July 2020 **Accepted**: 25 July 2020 **Publication**: 31 July 2020*

recursive feature elimination technique to select subset of features. Prediction model is derived using Auto Sklearn and neural network algorithms and achieved prediction accuracy of 72%.

Alvin R Mallcdem[11] developed models for the prediction of occurrence and severity of rice blast disease data collected from 2 government agencies from northern philippnes. In this study Authors have used principal component analysis (PCA) for selecting subset of features that is selection of most important features that influence in occurrence of disease and for predicting occurrence of blast disease ANN and Support Vector Machine (SVM) were used and to predict severity of blast disease regression model was used. Authors have verified models that were build significantly given good results.

Yangseon Kim[12] has proposed an artificial intelligence based prediction model for paddy blast disease. For this study authors have considered both historical data of blast disease occurrence and historical environmental data to build region specific models. Authors have considered 3 different regions like cheolwan, Icheon, Milyang of South Korea for the prediction of blast disease. In this study Long Term Memory Networks (LSTM) were used as predictive models .

Chethan B.S.[14] presented status of rice disease and measures followed. Authors has surveyed different regions of Karnataka district where rice cultivation is relatively high. And concluded that there is evidence in increase in incidence of major disease such as blast, shealth blight etc. severity of disease remained constant in few areas in other region it has increased due to cultivation practices that were followed by the farmers.

To brief out summary of related work authors have investigated weather parameters that influence the incidence of rice blast disease in different regions of the world. There is no research evidence that has investigated influence of weather parameters in disease incidence in Davangere district of Karnataka State. Few authors have developed predictive model for rice blast considering only weather factors and there is also issue of achieving of highest accuracy. To address these issue in proposed work along with investigating weather parameters we have also investingated disease related data provided Agricultural and Horticulture Research Station (AHRS) Kathalgere Davangere District. Further work is carried out to select subset of features using filter and wrapper approaches of feature selection.we have also trained ensemble classifier for achieving better accuracy and trained model also validated using K fold cross validation technique.

III. MATERIALS AND METHODS

Common Rice Disease that accur in all the stages of its growth is rice blast. Compare to other diseases rice blast has more compound data. And prediction can be done for both the season. Based on above parameters we decided to find relationship between rice blast and weather parameters and we have studied rice blast occurrence in previous years. In this section we have described Dataset that has been collected for the study and steps that are followed to predict rice blast at earliest.

A. Data Collection and Description

All India Co-ordinated Research Project on Integrated Farming Systems is operating at Agricultural and Horticultural Research Station (AHRS), Kathalagere, Tq: Channagiri, Dist: Davanagere. it is one of the AICRP project operating under the University of Agricultural and Horticultural Sciences, Shivamogga, District, Karnataka state and which is the Main Centre for Cropping System Research (MCCSR) since 1987-88 in the University (earlier at UAS, GKVK, Bangalore). Department of Agriculture and Horticulture Research Station Kathalgere provided the historiacal weather data parameters. This research station is situated between $13^{0}2^{1}$ N latitude and $76^{0}15^{1}$ E longitude at an elevation of 561.6m above the MSL.The historical weather data attributes are as follows:
The collected weather data last from 2013 to December 2019 and attributes of weather data are as follows:
Temperature Minimum: Minimum Temperature averaged for the week(in Celsius).
Temperature Maimum: Maximum Temperature averaged for the week(in celsius).
Temperature Difference: Difference of Maximum Temperature and Minimum Temperature.
Relative Humidity: Average Amount of Vapour in the air averaged for the week(in percentage.
Rainfall: the average precipitation count for the week (in millimeter)
Number of Rainy Days: if the rainfall in millimeter is greater than 2.5mm for the day then day is considered and as rainy day. Number of rainy days for the week is considered.
Wind speed: average air speed for the week (in kilometer)
Along with mentioned historical weather factors disease related data is also considered for making prediction more accurate. For collection of these data is supported by AHRS Kathalgere, Join Director office of Agriculture Davangere Distict and Agriculture website of government of Karnataka raitamitra.kar.gov.in and attributes of disease data are as follows:
Report Date of Blast Disease: the date of the cropping growth stage when rice blast has occurred.
Rice Crop Variety : Varities of seeds cropped by the farmers and some seeds are approved by the Seed Council government of Karnataka.
Severity of Rice Blast: Intensity of diseases with respect to growth stage.
Growth Stage: Different Stages of cropping rice are: Sowing, Tillering, Flower Intiation, Panicale Intiation, Maturity and Harvesting.
Season: Cropping Season of Rice are of 2 Season namely summer season from January to June and Kharif Season from July to December.

TABLE I: WEATHER PARAMETERS INFLUENCE BLAST DISEASE INCIDENCE

Date	Year	Minimum Temperature	Maximum Temperature	Temperature Difference	Relative Humidity	Rainfall	Number of Rainy Days	Wind Speed
Jan 1-7	2013	17	30	23.5	94	0	0	5.3
Jan8-14	2013	17	30	23.5	92	0	0	5.4
Jan 15-21	2013	17	31	24	92	0	0	5.5
Jan 22-28	2013	18	31	24.5	91	0	0	5.2
Jan 29- Feb 4	2013	19	32	25.5	92	0	0	5.3

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
Article Received: 15 July 2020 Revised: 20 July 2020 Accepted: 25 July 2020 Publication: 31 July 2020

Table I gives sample information related to weather parameters collected from southern transistion zone 7 kathalgere research station operating under University of Agricultural and Horticulture Science Shimoga. Data is available from Jan 2013 to December 2019. Table I defines different variables such Minimum Temperature, Maximum Temperature Temperature Difference in Celsius. Relative Humidity in Percentage Rainfall in Milli Meter. Wind Speed in Kilometer and Finall Number of Rainy Days if rainfall for the day is more than 2.5mm than day is considered as Rainy Day.

TABLE II: DISEASE DATASET

Stages	T1 Seed1	T2 Seed 2	T3 Seed3	Severity	Report Date	Season
Sowing	Sri Rama Sona	Jaya	Jyothi	0	0	Summer
Tillering	JGL 1798	Kaveri Sona	MTU1010	0	0	Kharif

Table II gives information related to paddy disease dataset collected from consulting Agriculture Scientists first variable in table is about Stages that defines 6 stages of paddy cultivation. T1 Seed1 is called Taluk 1 Varities of seeds cultivated from the year 2013-2019 similary T2 refers to Taluk 2 and T3 is called Taluk 3. 4th stage is severity if disease occurs in sowing stage then severity is 1 and that goes on. Report Date refers to Blast disease reported in previous years. Last Variable Season refers to two seasons of paddy cultivation.

B. Methodology Defined For Proposed Work

The approach is defined to build Prediction model for rice blast disease. We attained a relationship between weather parameters along with non weather parameters such as varities of seeds, growth stage, severity etc with disease or no disease data. According to the study investigated by plant Pathologist when temperature is low that is minimum temperature in connection with relative humidity is too high influence blast disease incidence.

According to the survey conducted by the agriculture experts during the year 2013 ,2014 and 2015 at different taluks of Davangere Distict of Karnataka State for the varities of seeds such as JGL1798 and Kaveri Sona there was an occurrence of rice blast disease with severity of 5-10% this data is also recorded in the study.

The Ensemble Techniques such as Bagging, Boosting and Stacking models are used for the prediction to achieve good accuracy . Ensemble Leaning is machine learning prototype where it combines multiple models these models are trained with independent variables such as weather related data and non weather related data and take disease and no disease data as dependent variables. Once forewarning model is confirmed to perform better than predecessor models. This can be used in the real field. Fig 1. Overiews the methodology used to predict rice blast disease.

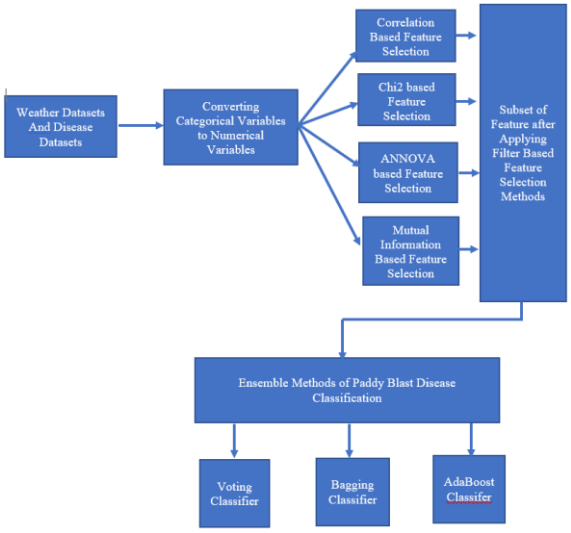

Fig 1: Methodology of the Proposed Work

IV. FEATURE SELECTION METHODS AND ENSEMBLE CLASSIFIERS

In this paper the models were trained using Weather Data and non weather data that is Blast disease data. Model is trained with 70% of total data and 20% as test data. It is made up of 364 instances. Different operations were performed on the blast disease dataset before the feature selection methods are applied. Two Operations performed are: Normalizing data and Preprocessing of data.

A. Preprocessing Phase using One-Hot Encoding

Data Set used for the classification and prediction as both numerical values and categorical data it is shown in Table I. Ensemble classifiers accepts only numerical data as input for the process of training and testing. Just to convert categorical values into numerical values a pre processing step is required One Hot Encoding is most extensive approach and it works well for converting categorical data to numerical data. This creates new columns(binary columns) indicating the presence of each possible value from the original data. Dataset represented in the Table I Feature 8,9,10,11 and 14 has categorical values . binary values are assigned to each variable to convert these features in the test and trained dataset. Eg: Sowing =1,Tillering-2 etc.

B. Min-Max Scaling

The tqo most important techniques for scaling numerical variables before modelling includes normalization and standardization. Normalization scales that from range of 0-1 seperately. Which will be in the range of floating values where we can have most precision. Normalization refers to the rescaling of data so that new values fal within the range of 0 and 1. A value is normalized as follows.

$$Y=(X-min)/(max-min) \qquad \text{Eq. (1)}$$

For our dataset we have variables that has different ranges to make variables fall within same range of 0-1 we have applied min max scaling method before we apply feature selection methods. Example is as follows

If min=14 and max=27 for the variable Minimum Temperature And X=17. If we apply these values to Eq. (1)

$$Y=(17-14)/(27-14)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
Article Received: 15 July 2020 Revised: 20 July 2020 Accepted: 25 July 2020 Publication: 31 July 2020

Y=3/13
Y=0.23

If we provide an X value that is out of bound of minimum and maximum value the resulting value y will not be in range of 0 and . if we observe above said values we have to remove those values that will not pertain to make predictions.

C. Feature Selection Methods

In recent years the magnification growth of Internet the many number of devices and tools for communication have created enmerous of data .To process and extract important data from this large data machine learning tools plays vital role. Applyying those tools with this enemrous amount of data with high dimensionality is time consuming and affects accuracy of the knowledge extracted. Redundant and irrelevant data affects performance of the model and unsupportable to memory requirements. To come out from other issues such as curse of dimensionality and to improve accuracy and performance of model another important preprocessing phase is needed. As we know preprocessing is used to remove noisy data but this can also be used to select, extract and combine extracted features.

Feature selection is a process that is more frequently used in machine learning and data mining applications. FS aims to reduce the curse of dimensionality by removing and irrelevant features to improve performance of the model.one way to improve accuracy of model by selecting subset of features which represents complete data. Irrelevant and redundant data are nothing but data with low dimensionality. A feature is considered as relevant if it integrated with the target class. Feature Selection as two important parameters to consider: Feature Search and Feature Evaluation.

b) Feature Search: Feature Search is a method that determines subset of features. Optimal solution can be found through exhaustic search. If we have n features then 2^n-1 possible feature subsets thus this very time consuming and impractical hence there are other many other feature search techniques that are feasible forward selection and backward elimination and genetic algorithms. If we have high dimesional data each feature will be evaluatued with some criteria that satisfies some conditions those features which satisifies some condition will be ranked top.

c) Feature Evaluation: Through this feature evaluation it is possible to evaluate relavant and redudancy in the features. Evaluation criteria has to be set and applied on each features because different criteria leads in selecting different set of features .

D. Filter, Wrapper and Embedded Based Feature Selection Methods

Filter based feature selection is independent of any machine learning techniques. Features in the data are selected on the basis of their score with the output variable.

In wrapper based approaches we try to select subset of features and train model using them. Based on the inferences drawn from the model we decide to add or remove features .

Embedded based feature selection is a combination of filter and wrapper approaches. This has its own feature selection methods that selects subset of features and selected features are used to train model. Figures and Tables.

In our proposed work we have used filter based feature selection methods like Pearson's Correlation and Chi S to select subset of features.

a) Pearson's Correlation: It is used to check dependency between two variables X and Y. The values varies from -1 to +1.

Correlation is given by:

$$\rho X,Y = \frac{COV\ (X,Y)}{\sigma X \sigma Y}$$

b) Chi Square: Chi Square is used in statistics to check independence of two variables x and y . for the given x and y we get observed output O and expected output E. and chi2 estimates how expected output E deviates from Observed output O.In feature selection we aim to select the features which are highly dependent on target variable. Higher chi2 indicates variable is highly dependent on target and can be considered to train model and lower chi2 value indicates variable is not dependent on target

c) Mutual Information: Mutual Information between two variables measures dependence of one variable with another variable. For example if X and Y are two variables

- If X and y are independent. Then no information can be obtained from X to Y. hence mutual information is equal to 0.
- If X is dependent on Y. then we can determine X from Y and Y from x with mutual information 1.
- When mutual information is 0<mutual information<1 then we can select features based on ranking method.

d) ANNOVA F Value: ANNOVA is called as analysis of variance and is a parametric statistical hypothesis test performed to test whether means from three to four samples of data comes from same distrubution or not. It calculates the ratio two variance. ANNOVA is a type of F statistic referred to as an ANNOVA f-test. It is used when we have to perform classification task. The results of this can be used in such a way that variables that are independent to target variable can be removed from classification task.

E. Ensemble Classifiers

Ensemble Classifiers are used to increase the results of machine learning. It combines multiple machine learning models together for doing similar task. Thus this technique improves accuracy of the predictive model than the single classifier used. In this we consider N learning models on a single training data to achieve N various models. Once different models are obtained the system combines their outputs in order to take final decision.

Ensemble methods used for prediction of Blast disease are:

a) Bagging
b) Adaboost
c) Stacking
d) Voting

A. Bagging

Bagging is also known as aggregation also known as bootstrap aggregation. Goal of Bagging is used to reduce the decision tree classifier variance . objective of this is to create subsets of data from training sample chosen randomly with replacement. Each subset of collection of data is used to train their decision

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
***Article Received**: 15 July 2020 **Revised**: 20 July 2020 **Accepted**: 25 July 2020 **Publication**: 31 July 2020*

trees. Average of all predictions from different trees are used which is more robust than single classifier.

B. Boosting

Boosting is a ensemble technique this generally decreases bias error and builds better predictive model. Boosting refers to a group of algorithms which converts weaker algorithms to stronger algorithms. Boosting involves multiple learners. Samples are weighted in each loop patterns that are misinterpreted are identified and new weights are assigned and weights of such patterns are also increased . thus the performance of ensembling process can be increased.

C. Stacking

In case of this ensembling technique in which multiple classifiers are combined via meta classifier. Layers are placed one after the other. Bottom layer takes the input from the original dataset and make the prediction and output of the bottom layer is again input to its top layer and thus it make the prediction based on the input it has received from its ouput layer. Thus accuracy and performance of the models can be increased efficiently.

D. Voting

Voting is one of the simple way of combining many prediction models from multiple machine learning algorithms. We can train data using different machine learning algorithms and ensemble them to predict the final output. Final output on a prediction is taken by majority vote. This is two strategies: hard voting and soft voting. Hard voting decides what ever class that receives highest number of votes will be considered. For example if we taken 3 classifier in which classifier 1 as predicted class 0, classifier 2 as predicted class 0 and classifier 3 as predicted class 1 on the basis of hard voting class 0 will be predicted as final. In case of soft voting predicted class are summed up or averaged.

V RESULTS AND DISCUSSIONS

Through the process investigated in the paper, a paddy blast prediction model is applicable to the southern transisition zone 7 Davangere district of Karnataka state with different input variables are developed. Our model is able to predict occurrence of blast disease based on environmental factors such as Minimum temperature, Maximum Temperature, Rainfall, Relative Humidity etc and non environmental factors called disease data such as varities of seeds, severity, report date etc.Before model is trained filter based feature selection methods are applied to select subset of features. 4 filter based features selection methods are applied in our study and top 15 dependent features with target variable are selected for further to tain the ensemble model. Fig. 2 and Fig.3 depicts ANNOVA f classify and Mutual information based feature selection methods that defines dependency of each and every variable with target variable.

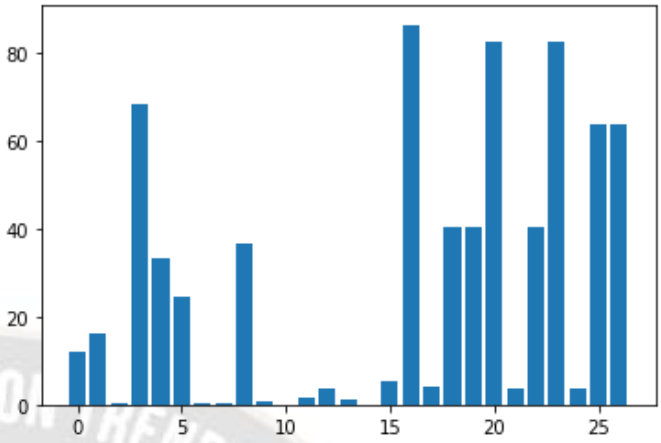

Fig 2: ANNOVA F_classify Feature Selection

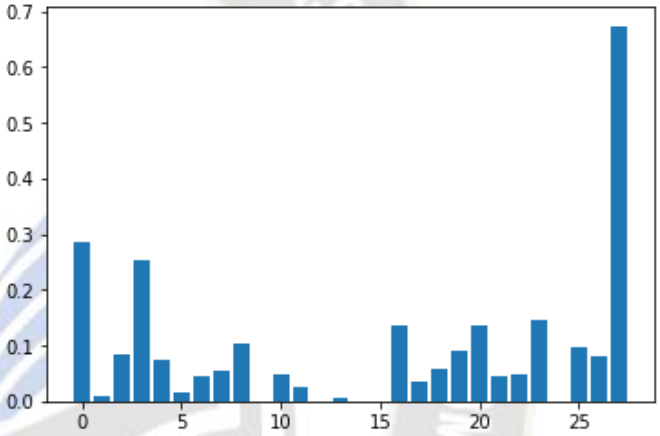

Fig 3: Mutual Information based Feature Selection

Table III depicts subset of features selected after applying 4 different filter based feature selection methods such as Pearson's correlation, ChiSquare, ANNOVA F Value and Mutual Information these features are used to train model to achieve good accuracy.

TABLE III: SUBSET OF FEATURES SELECTED TO TRAIN MODEL

Filter Based Feature Selection Applied	Features Selected Considering all feature selection methods to train model
Pearson's Correlation+ Chi Square+ ANNOVA F Value+ Mutual Information	Feature 0, Feature 1, Feature 3, feature 4, feature 5, feature 8, feature 16, feature 18, feature 19, feature 20, feature 22, feature 23, feature 25, feature 26

Table IV depicts classification report of different ensemble classifier. The report shows main classification metrics are accuracy, precision, recall and f1 score on a per class basis. The metrics are calculated based on true positive, true negative, false positive and false negative. Table IV also depicts precision recall score and ROC score.

Accuracy: $\frac{True\ Positive + True\ Negative}{Total\ Sample}$

Precision: $\frac{True\ Positives}{True\ Positives + False\ Positives}$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
Article Received: 15 July 2020 Revised: 20 July 2020 Accepted: 25 July 2020 Publication: 31 July 2020

Recall: $\frac{True\ Positives}{True\ Positives + False\ Negative}$

F1 Score:It is a Harmonic mean between precision and Recall the range of this is between 0 and 1. As defined greater F1 Score better the performance of model and mathematically this can be expressed as:

F1 Score: $2 * \frac{1}{\frac{1}{precision}+\frac{1}{Recall}}$

Precision-Recall Curve: This is used to measure of success of prediction. It is a tradeoff between precision and recall for different threshold. Greater area represents high recall and high precision.

Area Under Curve:It is used when we have a binary classification problem. Area under curve is a plot between false positive rate vs True positive rate.

TABLE IV: CLASSIFICATION REPORT OF DIFFERENT ENSEMBLE CLASSIFIERS

	Classification Report of Ensemble Classifiers					
	Accuracy	Precision	Recall	F1 Score	Precision Recall Score	ROC Score
Voting Classifier	0.88	0.86	0.88	0.87	0.81	0.88
Bagging	0.98	0.98	0.98	0.98	0.97	0.98
Adaboost	0.97	0.96	0.98	0.97	0.95	0.97

Fig. 4,6,8 depicts AUROC(Area Under Receiver Operating Characterstic). The ROC is plotted with TPR against the FPR where y axis is TPR and x axis is FPR.when model's AUC is 1 or near to 1 it has good measure of seperability. When model's AUC is 0 it has worst seperability. If AUC is 1 or near to 1 it clearly distinguish two classes. If AUC is 0 or near to 0 model overlaps solution that is it predicts positive as negative and vice versa. In our proposed solution voting classifier has 88% chances of predicting positive class as positive and negative class as negative. Similarly Bagging and Adaboost solutions has 98% and 97% respectively.

Fig. 5,7,9 gives information about precesion-recall curve it is a tradeoff between precision and recall for different threshold value. If the curve is very high then it simply explains that model has high precision and high recall. If both precision and recall remains high for different thresholds then models is giving accurate results and high positive results. If model has high recall but low precision then what ever model has predicted labels are incorrect. If model has high precision but low recall value then predicted lables are correct but they will be very few. Our proposed solution has precision recall score of voting classifer is 87% and 97%, 95% for Bagging and Adaboost classifiers this shows that all three different ensemble classifiers has high precision and high recall hence all 3 ensemble classifiers proposed to predict paddy disease is predicting labels correctly with high precision and high recall at different thresholds.

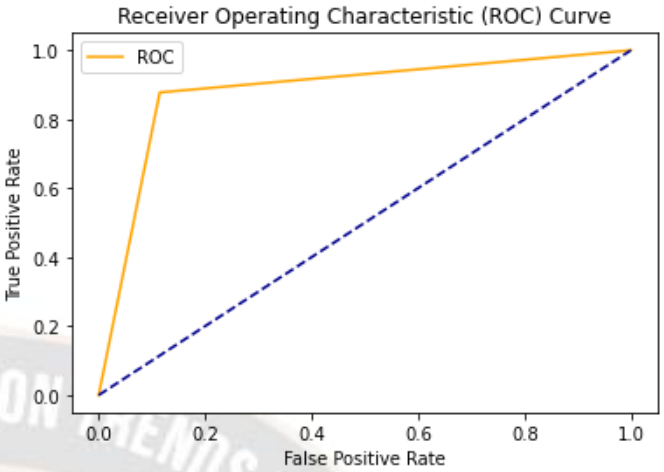

Fig. 4: ROC of Voting Classifier

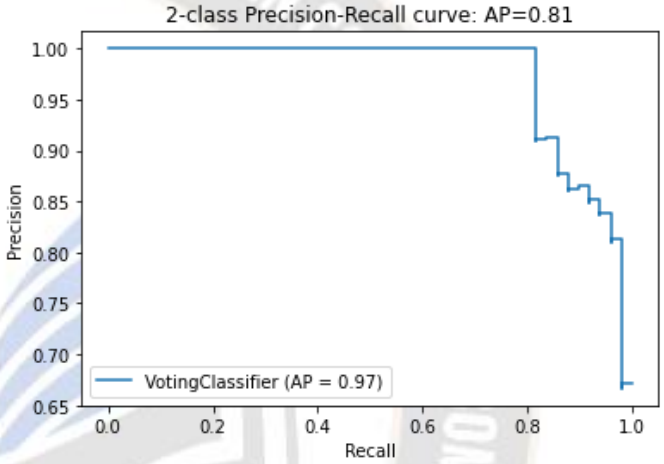

Fig.5: Precision-Recall curve of Voting Classifer

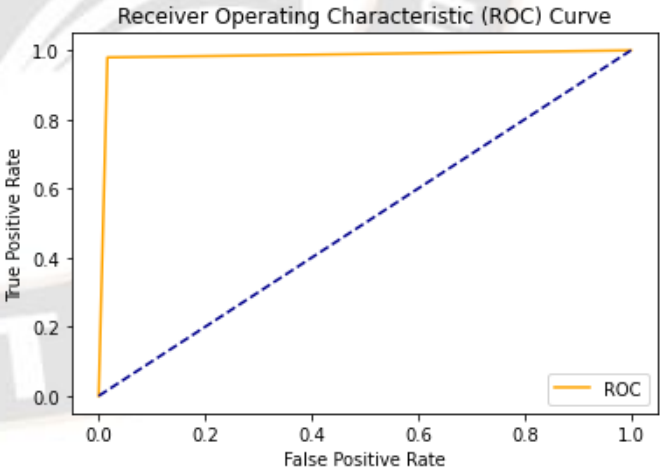

Fig. 6: ROC of Bagging Classifier

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
***Article Received**: 15 July 2020 **Revised**: 20 July 2020 **Accepted**: 25 July 2020 **Publication**: 31 July 2020*

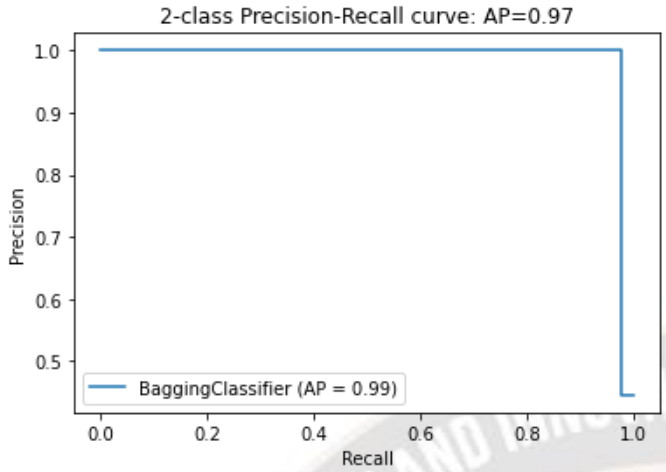

Fig.7: Precision-Recall Curve of Bagging Classifer

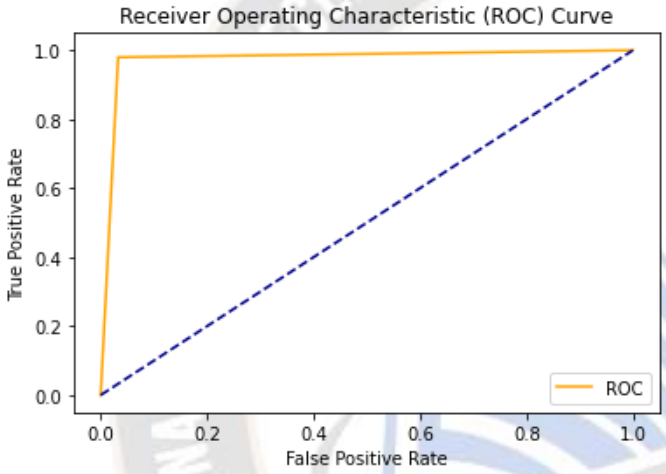

Fig.8: ROC of AdaBoost Classifer

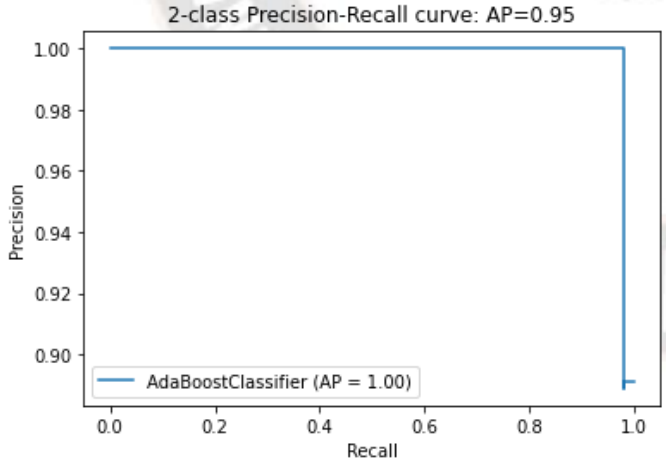

Fig. 9: Precision-Recall Curve of AdaBoost Classifer

*Conclusion:*In our proposed solution we have tried to solve the problem of paddy blast prediction by using machine learning approaches. In proposed solution we considered both disease data and weather datasets that influence in occurrence of blast disease. Filter based feature selection methods are used to select subset of features that improves the efficiency of classification model. According to the results obtained Minimum Temperature, Humidity, Rainfall has high influence in occurrence of disease and even non weather parameters such stages of cultivation influence disease to spread. Ensemble classifiers are used in our proposed model to train classifer and it is observed that all 3 classifiers that are trained such as Voting classifer, Bagging, Adaboost has achieved high accuracy with high precision and recall. To have a close look Bagging Classifier has achieved 98% of accuracy with 98% of precision and recall compared to other models bagging model has performed well in predicting paddy blast disease.

Acknowledgement: This Research is supported by Agricultural and Horticulture Research Station Kathalgere Davangere district running under Univeristy of Agriculture and Horticulture Science Shimoga.

References

[1] Shafaullah, Muhammad Aslam Khan, Nasir Ahmed Khan And Yasir Mahmood, “Effect Of Epidemiological Factors On The Incidence Of Paddy Blast (Pyricularia Oryzae) Disease,” Pak J.Phytopayhol.,Vol 23(2):108-111, 2011.

[2] Kwang-Hyung Kima, Jaepil Cho, Yong Hwan Lee, Woo Seop Lee, “Predicting potential epidemics of rice leaf blast and sheath blight in South Korea under the RCP 4.5 and RCP 8.5 climate change scenarios using a rice disease epidemiology model, EPIRICE,” Agricultural and Forest Meteorology 203(2015) 191 207.

[3] Yoshihiro Taguchi, Mohsen Mohamed Elsharkawy, Naglaa Hassan, Mitsuro Hyakumachi, “ A novel method for controlling rice blast disease using fan-forced wind on paddy fields”, Crop Protection 63(2014) 68-75.

[4] Kwang-Hyung Kim & Jaepil Cho, “Predicting potential epidemics of rice diseases in Korea using multi-model ensembles for assessment of climate change impacts with uncertainty information”, Climatic Change (2016) 134:327–339 DOI 10.1007/s10584-015-1503-2.

[5] G. Miah, M. Y. Rafii, M. R. Ismail, M. Sahebi, F. S. G. Hashemi,O. Yusuff and M. G. Usman, “Blast Disease Intimidation Towards Rice Cultivation: A Review Of Pathogen And Strategies To Control”, The Journal of Animal & Plant Sciences, 27(4): 2017, Page: 1058-1066 ISSN: 1018-7081.

[6] Rupankar Bhagawati, Kaushik Bhagawati†, D. Jini, R. A. Alone, R. Singh, A. Chandra, B. Makdoh, Amit Sen and Kshitiz K. Shukla, “Review on Climate Change and its Impact on Agriculture of Arunachal Pradesh in the Northeastern Himalayan Region of India”, Nature Environment and Pollution Technology An International Quarterly Scientific Journal p-ISSN: 0972-6268 e-ISSN: 2395-3454 Vol. 16 No. 2 pp. 535-539 2017.

[7] Rajendra Prasad, Anupam Sharma and Sweta Sehgal, “ Influence of weather parameters on occurrence of rice blast in mid hills of Himachal Pradesh”, Himachal Journal of Agricultural Research 41(2): 132-136 (2015).

[8] Rakesh Kaundal, Amar S Kapoor and Gajendra PS Raghava, “Machine learning techniques in disease

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 8 Issue: 7
DOI: https://doi.org/10.17762/ijritcc.v8i7.5435
***Article Received**: 15 July 2020 **Revised**: 20 July 2020 **Accepted**: 25 July 2020 **Publication**: 31 July 2020*

forecasting: a case study on rice blast prediction", BMC Bioinformatics 2006, 7:485 doi:10.1186/1471-2105-7-485.

[9] Y.H. Gu, S.J. Yoo, C.J. Park, Y.H. Kim, S.K. Park, J.S. Kim, J.H. Lim, "BLITE-SVR: New forecasting model for late blight on potato using support-vector regression", Computers and Electronics in Agriculture 130(2016) 169-176.

[10] Jia-You Hsieh, Wei Huang, Hsin-Tieh Yang, Chia-Chieh Lin, Yo-Chung Fan, Huan Chen, "Building the Rice Blast Disease Prediction Model based on Machine Learning and Neural Networks", EasyChair Preprint.

[11] Alvin R. Malicdem, Proceso L. Fernandez, "Rice Blast Disease Forecasting for Northern Philippines", Article in WSEAS Transactions on Information Science and Applications · January 2015.

[12] Yangseon Kim, Jae-Hwan Roh and Ha Young Kim, " Early Forecasting of Rice Blast Disease Using Long Short-Term Memory Recurrent Neural Networks", Sustainability 2018, 10, 34; doi:10.3390/su10010034.

[13] Dimitrios Katsantonis1 , Kalliopi Kadoglidou, Christos Dramalis And Pau Puigdollers, "Rice blast forecasting models and their practical value: a review", Phytopathologia Mediterranea (2017), 56, 2, 187−216 DOI: 10.14601/Phytopathol_Mediterr-18706.

[14] Chethana B.S., Deepak, C.A., Rajanna, M.P., Ramachandra, C. and Shivakumar, N., "Current Scenario Of Rice Diseases In Karnataka", I.J.S.N., VOL.7 (2) 2016: 405-412.

[15] P. B. Jawade, Dattatray Chaugule, Devashri Patil, and Hemendra Shinde, " Disease Prediction of Mango Crop Using Machine Learning and IoT", ICETE 2019, LAIS 3, pp. 254–260, 2020 , https://doi.org/10.1007/978-3-030-24322-7_33.

[16] Karim Foughali, Karim Fathallah, Ali Frihida, "Using Cloud IOT for disease prevention in precision agriculture", 9th International Conference on Ambient Systems, Networks and Technologies, ANT-2018 and the 8th International Conference on Sustainable Energy Information Technology, SEIT 2018, 8-11 May, 2018, Porto, Portugal.

[17] Sehan Kim, Meonghun Lee, Changsun Shin, "IoT-Based Strawberry Disease Prediction System for Smart Farming" , Sensors 2018, 18, 4051; doi:10.3390/s18114051.

[18] S. Ramesh and Bharghava Rajaram, " Iot Based Crop Disease Identification System Using Optimization Techniques", ARPN Journal of Engineering and Applied Sciences, ISSN 1819-6608 VOL. 13, NO. 4, FEBRUARY 2018.

[19] Jennifer G. Dy, Carla E. Brodley, "Feature Selection for Unsupervised Learning", Journal of Machine Learning Research 5 (2004) 845–889.

[20] Shadi Aljawarneh, Muneer Bani Yassein, Monther Aldwair, " Anomaly-based intrusion detection system through feature selection analysis and building hybrid efficient model", Journal of Computational Science 25(2018) 152-160.