



e-ISSN: 2319-8753 | p-ISSN: 2347-6710

# IJIRSET

International Journal of Innovative Research in  
**SCIENCE | ENGINEERING | TECHNOLOGY**

# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 11, Issue 3, March 2022

**ISSN** INTERNATIONAL  
STANDARD  
SERIAL  
NUMBER  
INDIA

Impact Factor: 7.569

9940 572 462

6381 907 438

[ijirset@gmail.com](mailto:ijirset@gmail.com)

[www.ijirset.com](http://www.ijirset.com)



# Survey on Prediction of Land Suitability for Crop Cultivation Using Random Forest Machine Learning Algorithm

Mr. Pratap Bhan Singh Sarjan Singh Senger, Prof. Sushma Shinde

Department of Computer Engineering, Siddhant College of Engineering, Sudumbre, Pune, India

**ABSTRACT:** Crop yield prediction is critical in agriculture. Crop yield depends on various features including geographic, climate and biological. The major research on agricultural management is the development of cost effective methods for predicting the crop yield with the available parameters like irrigation details, fertilizer details and temperature. This research article discusses five Feature Selection (FS) algorithms namely Sequential Forward FS, Sequential Backward Elimination FS, Correlation based FS, Random Forest Variable Importance and the Variance Inflation Factor algorithm for feature selection. Machine learning techniques are generally well-suited for a particular region, thus assisting farmers a great deal in crop prediction. This paper offers a novel FS approach termed modified recursive feature elimination (MRFE) for crop prediction that selects acceptable features from a data set. Using a ranking algorithm, the suggested MRFE technique chooses and ranks salient features.

**KEYWORDS:** Feature Selection Algorithm, Machine Learning, Performance Metrics, Crop prediction

## I. INTRODUCTION

Crop prediction in agriculture remains a challenge, despite recent advancements that include the use of a variety of technology resources, techniques, and procedures. The development of cost-effective algorithms for predicting crop production with available characteristics such as irrigation information, fertilizer details, and temperature is a key focus of agricultural management research. Because the parameters differ for each zone, resulting in a large crop prediction data set, it is necessary to pick critical traits that aid in the identification of acceptable crops for certain land areas. Feature selection approaches are used in the process.

Crop yield can be predicted using historical data of crop production and weather data. A crop production dataset may contain parameters like year, area, production, and yield. A weather dataset may contain parameters like minimum temperature, maximum temperature, average temperature, precipitation, evapotranspiration, and reference crop evapotranspiration. These are the basic parameters of weather dataset for crop yield prediction; weather dataset may also contain more parameters.

The feature selection algorithms to identify relevant paddy field circumstances (features) in order to provide a comprehensive view of paddy crop output. The data analysts are working towards developing predictive models in the form of expert systems to improvise agricultural yield, considering the environmental features, soil quality, irrigation and land usage. Machine learning (ML) is used in most existing systems to estimate crop yields, but little has been done to predict region-specific crops based on soil and environmental factors.

To predict an appropriate crop and evaluate the effectiveness of the FS process, the features collected are supplied to the ML classifiers. A single prediction model (such as the kNN, NB, DT, and SVM as well as an ensemble prediction model and Bagging approaches, have been utilized to classify crop prediction in much of the study. Each algorithm has its own set of prediction features. However, a classifier that works well with the suggested FS approach for crop prediction must be found. The feature elimination (MRFE) technique is used for selecting the most appropriate key features from a permutation crop data set depending on soil and environmental conditions. The algorithm does not need to be updated with the data set at each iteration, so it reduces the computational time. MRFE's Performance Compared to Other FS Techniques Other Benchmark Data Sets and the Bagging Classifier: Miscellaneous data sets were retrieved from the UCI Repository to ensure that the suggested MRFE approach was applicable to data sets other than crop-related data sets.



## II. LITERATURE SURVEY

**M Gopal et al:** In this paper, five feature selection techniques are used to pick the features: sequential forward feature selection, sequential backward feature removal, correlation based feature selection, random forest Variable Importance, and Variance Inflation Factor. Important features are selected by feature selection algorithms depending on their selection criteria. The performance of feature selection algorithms is evaluated using the RMSE, MAE, R, and RRMSE metrics. MLR, ANN, and M5P are given the results of the feature selection. The forward feature selection approach is good for better prediction, and the best features for the paddy dataset for specified research region are Area, Open Wells, Tanks, Temperature Maximum, and Canal Length.

**Dipika H. Zala et al:** For agricultural yield prediction, we present an effective Bootstrap aggregating technique. The accuracy of crop yield prediction will be improved by using a voting cast of bagging technique with different parameters. The application of crop yield prediction is an important concern in data mining. It can be easily decidable that which climate or weather is suitable for crop.

**Bahl et al:** we looked at the suitability of an unsupervised strategy based on PCA and kNN, as well as a supervised approach based on RFs with and without previous feature selection, for predicting NM toxicity. The reduced RF model following backward RFE had the best performance in terms of balanced accuracy of the prediction model. In every example, variable selection based on the MDA produced equal or better outcomes than Gini significance. In both the full and reduced models following RFE, the three most critical parameters zeta potential, redox potential, and dissolution rate were among the highest ranking variables in both unsupervised and supervised analyses.

**P. S. Maya Gopal et al:** In this paper, The Bourta algorithm was used to select the most essential features for crop yield prediction. Crop area, canal length, number of open wells, number of tube wells, number of tanks, maximum temperature, average temperature, nitrogen, phosphorus, and potash fertilisers, sun radiation, and seed rate have all been identified as essential factors. These characteristics are fed into the MLR model to determine accuracy. By incorporating these features into the model, it was able to reach an accuracy of 84%.

**K.Rajini et al:** In this paper, By merging the IB1 learner and the A1DE updatable learner, a novel ensemble of heterogeneous incremental classifiers has been presented. The methodology was created primarily to accurately predict ART outcomes. For the ART data set used in the study, the suggested model accomplishes this goal with a ROC area of 94.1. Several ART data sets and several test choices are used to evaluate the model's efficacy. For the same data set, the proposed ensemble outperforms the others. To see if the suggested model is generalizable to other domains, it is compared to additional benchmark data sets and shown to perform better.

**J.-Y. Hsieh et al:** In this Paper, we tried to solve the prediction problem by using the Machine Learning as well as Neural Network methods. The classification model was built based on the climate and historical data. This classifier's projected results have a 72 percent accuracy rate. Increasing the number of microclimate parameters can improve prediction accuracy, according to this study.

**J. Camargo et al:** In this work, the importance of feature selection in raising accuracy levels to above 95% was a crucial conclusion of this study. On average, accuracy scores were 3.8 percent lower without feature removal. Feature selection offers the added benefit of allowing for faster real-time computations with fewer features. Chow-Liu trees were discovered to be a feature selection method that allows for the rapid selection of a strong collection of features with a minimal number of iterations and comparable accuracy to that of a forward selection strategy.

**R.RajashakerPullanagari et al:** In this paper, the viability of employing RF-RFE to combine hyperspectral, topography, and soil data to obtain pasture quality parameters of heterogeneous pasture was investigated in this work. Because many environmental and management factors influence pasture quality, our findings revealed that combining hyperspectral data with commonly accessible environmental characteristics (elevation, slope angle, and soil type) enhanced prediction accuracy when compared to hyperspectral data alone. This finding also revealed that by selecting the most sensitive factors throughout the spectrum and environmental data, RF-RFE significantly improved estimations of pasture quality (RPD = 2.11–2.35). Elevation, slope, and soil type were found to be key determinants in predicting CP, while the same variables were found to be significant in predicting ME, with the exception of elevation.

**F. Balducci et al:** The research presented in this paper introduces practical, low-cost, and simple-to-implement tasks that can help an agricultural company increase productivity while also deepening the study of the smart farm model;





technological progress in a field that requires control and optimization can truly contribute to conserving natural resources, adhering to business and international laws, meeting consumer needs, and pursuing economic profits. Machine learning and more traditional statistical techniques were used to utilize the three separate data sources, with a special focus on the IoT sensors dataset. The first task demonstrates that a neural network model can forecast total apple and pear crops on the Istat dataset with a success rate of nearly 90%, while the second task demonstrates that polynomial predictive and regression models are better suited for the CNR scientific data due to the nature of the dataset.

**M. Langoet al:** Two sorts of methodological contributions are presented in our work. To begin, we propose incorporating a random set of attributes into this ensemble. This approach enhances G-mean and sensitivity measurements for greater dimensional complex datasets, according to experiments. We've also noticed that, unlike the original RBBag, it increases the diversity of component classifiers and that employing a larger number of components improves classification. Second, for several imbalanced classes, we have presented an extension of Roughly Balanced Bagging that uses the multinomial distribution to estimate cardinalities of class examples in bootstrap samples. The undersampling version of our proposed Multiclass RBBag enhances G-mean and is better than the oversampling variation and simpler multi-class classifiers, according to testing with synthetic and real datasets

### Open Issues

- In existing System, the RFE method is a wrapper-type FS method that searches for a subset of features, starting with all features in the training data set and successfully deleting features until only a small number remain.
- The RFE technique ranks acceptable traits in order of relevance, rejecting the ones that aren't as vital.
- This method needs an iterative process for data set updating in the feature elimination process.
- The most challenging aspect of the RFE is updating the data set, and the greatest time is spent removing weak features.

### III. CONCLUSION

In this paper, the suggested MRFE approach is applicable to both crop and non-crop data sets. When compared to previous strategies, the MRFE uses permutation and ranking to choose the most suited characteristics with improved prediction ACC in the shortest amount of time. MRFE technique in predicting the most suitable crops for cultivation using classification algorithms.

### REFERENCES

1. M Gopal P S and B. R, "Selection of important features for optimizing crop yield prediction," *Int. J. Agricult. Environ. Inf. Syst.*, vol. 10, no. 3, pp. 54–71, Jul. 2019.
2. D. H. Zala and M. B. Chaudhri, "Review on use of BAGGING technique in agriculture crop yield prediction," *Int. J. Sci. Res. Develop.*, vol. 6, no. 8, pp. 675–677, 2018.
3. Bahl et al., "Recursive feature elimination in random forest classification supports nanomaterial grouping," *NanoImpact*, vol. 15, Mar. 2019, Art. no. 100179.
4. P. S. Maya Gopal and R. Bhargavi, "Feature selection for yield prediction in boruta algorithm," *Int. J. Pure Appl. Math.*, vol. 118, no. 22, pp. 139–144, 2018.
5. K. Ranjini, A. Suruliandi, and S. P. Raja, "An ensemble of heterogeneous incremental classifiers for assisted reproductive technology outcome prediction," *IEEE Trans. Comput. Social Syst.* early access, Nov. 3, 2020, doi: 10.1109/TCSS.2020.3032640
6. J.-Y. Hsieh, W. Huang, H.-T. Yang, C.-C. Lin, Y.-C. Fan, and H. Chen, "Building the rice blast Disease Prediction Model based on Machine Learning and Neural Networks," *Easy Chair World Sci.*, vol. 1197, pp. 1–8, Dec. 2019.
7. J. Camargo and A. Young, "Feature selection and non-linear classifiers: Effects on simultaneous motion recognition in upper limb," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 27, no. 4, pp. 743–750, Apr. 2019.
8. R. Rajashekar Pullanagari, G. Kereszturi, and I. Yule, "Integrating airborne hyperspectral, topographic, and soil data for estimating pasture quality using recursive feature elimination with random forest regression," *Remote Sens.*, vol. 10, no. 7, pp. 1117–1130, 2018.
9. F. Balducci, D. Impedovo, and G. Pirlo, "Machine learning applications on agricultural datasets for smart farm enhancement," *Machine*, vol. 6, no. 3, pp. 38–59, 2018.
10. M. Lango and J. Stefanowski, "Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data," *J. Intell. Inf. Syst.*, vol. 50, no. 1, pp. 97–127, 2018



**Impact Factor:  
7.569**



# INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

 9940 572 462  6381 907 438  [ijirset@gmail.com](mailto:ijirset@gmail.com)



[www.ijirset.com](http://www.ijirset.com)

Scan to save the contact details