A Survey on Crop Recommendation Using Data Mining Methods

S. Kiruthika* & Dr.D. Karthika**

*Assistant Professor, Department of Computer Science, P.K.R Arts College for Women, Gobi, Tamilnadu, INDIA.
E-Mail: kirthi.deepak{at}gmail{dot}com
**Associate Professor, Department of Computer Science, VET Institute of Arts and Science College, Erode, Tamilnadu, INDIA.

***Abstract*—**Data mining is the process of analyzing and extracting useful information from large amounts of data. Data mining is used in a variety of industries, including banking, retail, medical, and agriculture. In agriculture, data mining is utilized to analyze numerous biotic parameters. Agriculture is a major source of income and employment in India. The most prevalent difficulty faced by Indian farmers is that they do not select the appropriate crop for their land. They will experience a significant drop in production as a result of this. Precision agriculture has been used to solve the farmers' issue. Precision agriculture is a modern farming method that utilizes research data on soil characteristics, soil types, and crop yield data to recommend the best crop to farmers depending on site-specific parameters such as temperature, rainfall, and the farm's latitude, longitude, altitude, and distance from sea. This decreases the number of times a crop is chosen incorrectly and increases production. Nevertheless, RS-based systems need the processing of massive volumes of remotely sensed data from many platforms, thus automated ways to provide reliable recommendations are presently receiving more attention. This is owing to automation-based systems' capacity to manage a high number of inputs and non-linear activities. The article provides a survey of the various methodologies followed with the view of developing an automated system for the recommendation of crop in agriculture system.

***Keywords*—**Crop Recommendation; Farming Systems; Environmental Impact; Expert Knowledge; Precision Agriculture; Sensed Data; Yield Production.

***Abbreviations*—**Chi-square Automatic Interaction Detector (CHAID); K–Nearest Neighbour (KNN); Linear Support Vector Machine (LSVM); Neural Network (NN); Precision Agriculture (PA); Random Forest (RF); Synthetic Minority Over-sampling Technique (SMOTE).

I. INTRODUCTION

INDIA is one of the world's oldest countries with a flourishing agricultural sector. However, owing to globalization, agricultural practices have significantly changed in recent years. The state of agriculture in India has been influenced by a number of variables. To reclaim one's health, several innovative technologies have been developed. Precision agriculture is such an approach. In India, precision agriculture is growing rapidly. Precision agriculture is a type of "site-specific" farming technology. It has given us the benefit of more efficient input, output, and better agricultural decisions. Despite the fact that precision agriculture has improved, there are still certain difficulties. Many methods exist that suggest inputs for a specific piece of farmland. Crops, fertilizers, and even farming methods are suggested by systems [Reashma & Pillai, 1].

Crop recommendation is one of the most important aspects of precision agriculture. Crop recommendations are based on a number of factors. Precision agriculture seeks to discover these factors on a site-by-site basis in to overcome crop selection difficulties. Although the "site-specific" method has enhanced outcomes, there is still a need to monitor the systems' outcomes. Precision agricultural systems aren't all created equal. However, in agriculture, it is critical that the advice given are accurate and precise, as errors can result in significant material and financial loss. Many studies are being conducted in attempt to develop an accurate and efficient crop prediction model [Patrício & Rieder, 2; Chetan et al., 3].

However, RS-based systems need the processing of massive volumes of remotely sensed data from many platforms, thus automated ways to provide reliable recommendations are presently receiving more attention. This work will survey the different methods for agriculture crop recommendation and also will review the merits and demerits of that system. The conclusion of this study is that rapid technological advancements will give cost-effective and comprehensive solutions for better crop and environmental status estimates and decision making. Precision Agriculture (PA) is anticipated to include more focused integration of

diverse specialist knowledge and the creation of hybrid systems integrating multiple approaches in the near future [Hegde et al., 4; Wankhede, 5].

In this research work, Section II presents a literature review of crop recommendation in an improved agriculture system, Section III provides the overview on the classical approaches that are employed for crop recommendation, Section IV gives the comparison of the research strategies along with its benefits and drawbacks, and Section V discusses the results of simulation.

II. LITERATURE REVIEW

It present a discussion on the review details of crop recommendation models using different methods.

Kulkarni et al., [6] created a crop recommender that employs machine learning's ensembling approach. The ensembling approach is being used to create a model that integrates the predictions of many machine learning models in order to accurately propose the best crop depending on the soil type and features. Random Forest, Naive Bayes, and Linear Support Vector Machine are the independent base learners utilized in the ensemble model (LSVM). Every classifier generates a set of class labels that are accurate enough. The majority vote approach is used to merge the class labels of separate base learners. The crop recommendation algorithm divides the input soil information into Kharif and Rabi recommendable crop types. The dataset includes physical and chemical properties of the soil, as well as climatic conditions including average rainfall and surface temperature samples. Integrating the independent base learners yielded a classification accuracy of 99.91 %.

Pudumalar et al., [7] developed an ensemble model with majority voting approach to suggest a crop for site specific parameters with good accuracy and efficiency, utilizing Random tree, Chi-square Automatic Interaction Detector (CHAID), K-Nearest Neighbor, and Naive Bayes as learners.

Kumar et al., [8] developed a unique approach for predicting the best crop for a farmer, detecting pests that may harm the crop, and recommending pest control methods. In this study, the SVM classifier, the Decision Tree method, and the Logistic Regression method were used, and it was discovered that the SVM classifier provides greater accuracy than other techniques.

Cai et al., [9] used the United States Department of Agriculture's (USDA) Common Land Units (CLUs) to combine spectral data for every field based on a time-series Landsat image data stack to primarily resolve cloud contamination while employing a machine learning model based on Deep Neural Network (DNN) and high-performance computing for intelligent and scalable computation of classification processes. Tests were conducted to investigate which information is most valuable for training a machine learning model for crop-type classification, as well as how different geographical and temporal aspects impact crop-type classification performance, to obtain timely crop type information. The incorporation of temporal phenology information and uniformly distributed spatial training samples in the research domain enhances classification performance, according to computational studies.

Shakil Ahamed et al., [10] focused on using data mining approach to detect information from agricultural data in order to estimate crop yields for key cereal crops in Bangladesh's major districts.

Utilizing Long-Short Term Memory (LSTM), Neural Networks, satellite imagery, and weather data, Schwalbert et al., [11] proposes a new model for performing in-season ("near real-time") soybean production forecasts in southern Brazil. The aims of this project were to: 1) contrast the performance of three different methodologies (extension of linear regression, random forest, LSTM neural networks) for forecasting soybean yield using Normalized Difference Vegetation Index (NDVI), Enhanced Vegetation Index (EVI), land surface temperature, and precipitation as independent variables, 2) determine how early (during the soybean growing season) this technique is able to forecast with reasonable accuracy. Satellite and weather data were covered with a non-crop-specific layer based on field boundaries acquired from Brazil's Rural Environment Registry, which is required of all farmers. The report's main findings were: 1) soybean yield forecasts at municipality-scale with a Mean Absolute Error (MAE) of 0.24 Mg ha1 at DOY 64 (march 5) 2) For all forecast dates except DOY 16, when multivariate OLS linear regression produced the highest results, the LSTM neural networks outperformed the other techniques, and 3) model performance (e.g., MAE) for yield forecast worsened when forecasts were made earlier in the season, with MAE rising from 0.24 Mg ha1 to 0.42 Mg ha1 (latest values from OLS regression) when forecast timing was adjusted from DOY 64 (March 5) to DOY 16 (April 15). (January 6). This study shows how field survey data may be combined with statistical methods, remote sensing, and weather to provide more accurate in-season soybean production forecasts.

By examining trends in historical data, Raja et al., [12] attempted to forecast crop production and cost that a farmer may receive from his property. Using a sliding window non-linear regression approach to forecast based on many parameters influencing agricultural productivity, like rainfall, temperature, market prices, land area, and previous crop yield. The study is carried out for a number of districts in the State of Tamilnadu, India. The system will recommend the optimal crop selections for a farmer to respond to the current social crises that many farmers are facing.

Sadia et al., [13] suggested a recommendation method to assist farmers in determining which crops are most suited to a given soil. The suggested method uses a mobile application to conduct Pearson correlation similarity calculations for defining particular locations based on the user's soil type and geological information. Following that, a suggestion will be made based on the fruit production rate of the chosen locations. As a result, the approach will reduce erroneous fruit crop selection and hence enhance production. Using the

Precision and Recall technique, the created system is verified against real-world data and shown to be accurate.

By modifying the cost and gamma factors, Suchithra & Pai [14] developed a modified sigmoid kernel SVM classifier with greater performance. The research is being done for a paddy field multiclass soil fertilizer recommendation system. To modify the SVM parameters, several optimization approaches such as Genetic Algorithm and Particle Swarm Optimization are utilized. Lastly, a comparison of performance is performed for the various parameter selections, highlighting their accuracies.

Goldstein et al., [15] used several regression and classification methods on a dataset to create models that could predict the agronomist's suggested weekly irrigation plan. To discover which factors consistently contributed to prediction accuracy, the models were constructed using eight distinct subsets of variables. The top regression model was Gradient Boosted Regression Trees, with accuracy of 93 %, while the best classification model was the Boosted Tree Classifier, with accuracy of 95 %, according to the results (on the test-set). Data that did not contribute to the model's success rate prediction were also identified. The resultant model can greatly simplify the irrigation planning procedure for agronomists.

By using the Synthetic Minority Over-sampling Technique (SMOTE), Liu et al., [16] presented a clustering centers optimized approach. The procedure begins by examining the original sample points and selecting density-based grouping center. The clustering centers is then used to produce minority samples in order to balance out the data distribution. Lastly, for accurate prediction, the ensemble method is utilized to train the prediction model. Over unbalanced soil data, empirical findings demonstrate that suggested technique outperforms existing modern prediction algorithms in terms of prediction accuracy.

Kale & Patil [17] suggested a Marathi calendar based on nakshatras to help farmers decide which crops to grow. Its goal is to develop techniques that would help farmers improve their economic circumstances by allowing them to make educated decisions. The system's technique, in particular, employs data mining to create expert recommendations, as well as fuzzy logic and machine learning to provide suitable decisions to farmers for the growing of anticipated crops.

Kamatchi & Parvathi [18] provided a predictive analysis to determine the best crop that can be grown under particular weather circumstances, as well as a hybrid recommender system that uses CBR - Case-Based Reasoning to improve the system's success ratio. The collaborative filtering method and case-based reasoning are combined in this unique hybrid system. The model's unique feature is the use of district-level agricultural data analysis to forecast future climatic conditions and propose crops depending on those circumstances, as well as taking into account the district's agriculture pattern employing a hybrid recommender system.

Support Vector Machines (SVM) were proposed by Löw et al., [19] for crop categorization in irrigated landscapes at the object level. 71 multi-seasonal spectral and geo statistical characteristics derived from Rapid Eye time series are used as input to the classifications. A subset of features with the highest accuracies was chosen using the Random Forest (RF) feature significance score. Using two measures of uncertainty, the maximum a posteriori probability and the alpha quadratic entropy, the relationship between hard result accuracy and soft output from the SVM is examined. In particular, the soft outputs of the SVM, as well as traditional accuracy measures, are used to explore the influence of feature selection on map uncertainty. Consequently, the SVMs applied to the selected features subspaces that were constituted of the most informative multi-seasonal characteristics resulted in a substantial decrease in thematic uncertainty and a noticeable improvement in classification accuracy of up to 4.3 %. The size of the feature space has been proven to influence SVM, while RF-based feature selection has been found to help. SVM-based uncertainty metrics provide useful information on the spatial distribution of error in crop maps.

Prabakaran et al., [20] set out to estimate future productivity using data compiled by field specialists, as well as productivity influencing variables. The integration of fuzzy logic and Support Vector Machine is used to detect this feature. The suggested method has been fully tested on the ground, and the planned structural findings have shown significant benefits over the lack of a correct method. With a prediction accuracy of 95%, this method designed to adjust for performance decline produced greater productivity. Moreover, the suggested intelligent integrated decision support system offered a deficiency level of needed input scale to improve production and reduce fertilizers use in agriculture. To create such a method to manage fertilizers use, a 30-year climatic parameter was taken into account.

Table 1: Existing Work Inferences

Researcher	Technique	Advantages	Limitations
Löw et al., [19]	Support Vector Machines	Better performance	It does not suitable for large data sets
Shakil Ahamed et al., [10]	Data mining techniques	Higher accuracy	Computational complexity is high
Pudumalar et al., [7]	K-NN and Naive Bayes	Effective and accurate	Does not used a data set with huge volume of attributes
Raja et al., [12]	Non-linear regression technique	Provides higher accuracy	Does not tested with real time application
Kulkarni et al., [6]	Ensembling technique of machine learning	Produces better accuracy	Time consuming nature

Cai et al., [9]	DNN	Gives better classification accuracy	Computational complexity is very high.
Suchithra & Pai [14]	Sigmoid kernel SVM classifier	Improved recommendation	It does not suitable for large data sets
Kulkarni et al., [6]	Ensemble RF, NB, and Linear SVM	Produces higher accuracy	Time consuming nature
Goldstein et al., [15]	Different regression and classification algorithms	Higher accuracy	It takes long time for training.
Kumar et al., [8]	SVM classifier, Decision Tree method and Logistic Regression method	Higher accuracy	Very expensive
Kale & Patil [17]	fuzzy logic	Provides better performance	Produces approximate results
Schwalbert et al., [11]	LSTM	More reliable	LSTMs take longer time for training
Liu et al., [16]	Clustering center optimized method by Synthetic Minority Over-sampling Technique (SMOTE)	Flexible and effective	Increase the overlapping of classes
Prabakaran et al., [20]	Fuzzy logic and support vector machine	Achieved higher productivity	It produces very approximate results
Sadia et al., [13]	Pearson correlation	Flexible	Lesser accuracy

III. Inferences from the Existing Works

The profitability of a crop is mostly determined by weather conditions, crop output, and cultivation and production expenses. The farmer's economic well-being is determined not only by the crop's production, but also by the demand for the commodity. Because agriculture is India's major source of income, several variables must be considered while choosing a crop because it affects the farmer's financial well-being. It would benefit farmers' economic wellbeing if a farmer could be advised if a certain crop will be profitable or not based on important criteria such as yield, weather prediction, market demand, and a few additional expenses. Traditionally, the crop's profitability is only recognized towards the conclusion of the harvest. Using data-driven Machine Learning (ML) models using historical data to predict whether a crop will be profitable or not based on current cost circumstances, weather prediction, crop yield, and minimum predicted sale price might assist predict whether a crop will be profitable or not. Recently used many machine learning methods for prediction and classification to suggest recommendation possesses the minimum accuracy due to the inability of choosing the appropriate features. It requires high processing time.

IV. Solution

India is known for its diversity, which includes a wide range of physical and cultural conditions. Almost every family in India is reliant on agriculture and agricultural-related occupations. Precision agriculture is not highly valued in India. To provide an enhanced recommendation, further research can propose a preprocessing model to get proper input data like scale variation so that, can get accurate results. And then introduce an optimization based significant feature selection model to reduce the computation time and complexities and finally will use an enhanced deep learning model to solve accuracy related problem for further research.

V. Results and Discussion

In this research work, Section II presents a literature review of crop recommendation in an improved agriculture system, Section III provides the overview on the classical approaches that are employed for crop recommendation, Section IV gives the comparison of the research strategies along with its benefits and drawbacks, and Section V discusses the results of simulation.

5.1. Experimentation Metrics

Recall is defined as ratio of relevant instances, which is obtained over the overall number of relevant cases. Both precision and recall are hence dependent on an understanding and the measure of relevance. Greater recall shows that method retrieved many of relevant outcomes. Mathematically, recall is defined to be as below:

$$Recall = \frac{TP}{TP+FN} \quad (1)$$

The proportion of relevant instances among retrieved instances is denoted as precision. When an algorithm returns a lot more relevant results than irrelevant ones, it's said to have high precision. Precision is defined as below:

$$Precision = \frac{TP}{TP+FP} \quad (2)$$

Accuracy is one metric used for the evaluation of classification models. In layman's terms, accuracy is the percentage of predicted values that are correct. The accuracy of binary classification may also be calculated in terms of positives and negatives, as shown below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

Here, TP - True Positive, TN - True Negative, FP - False Positive, FN - False Negative.

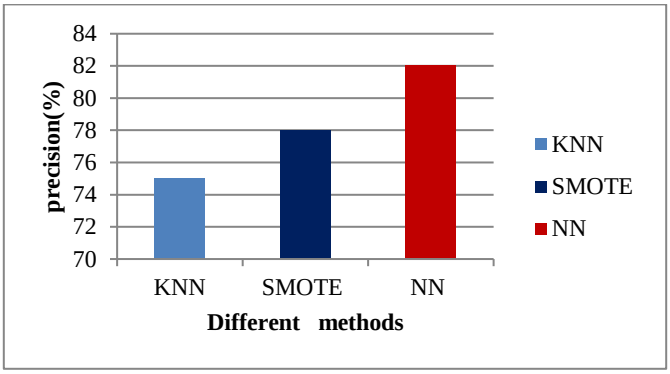

Figure 1: Precision Outcomes against Classifiers

Accuracy performance measure comparison among suggested NN and current KNN, SMOTE as indicated in figure.1. it indicates suggested method's effectiveness. By the outcomes it is concluded that suggested NN method produces higher precision results of 82% while the existing KNN, SMOTE models produces only 75% and 78% accordingly.

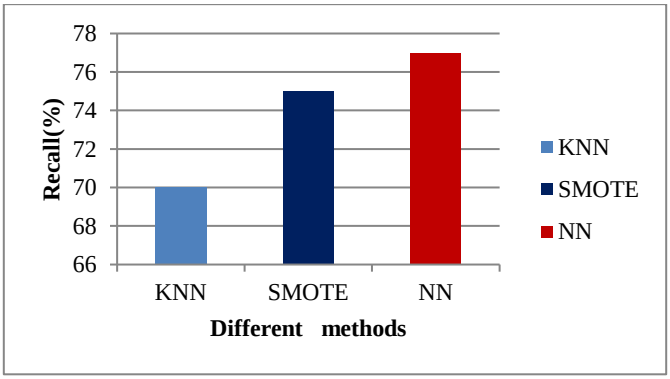

Figure 2: Recall Outcomes against Classifiers

Figure: 2. indicates performance measures comparison among suggested NN and current KNN, SMOTE in terms of recall and it shows the proposed techniques effectiveness. Through outcomes it is concluded that suggested NN model produces higher recall results of 77% while the existing KNN, SMOTE models produces only 70% and 75% accordingly.

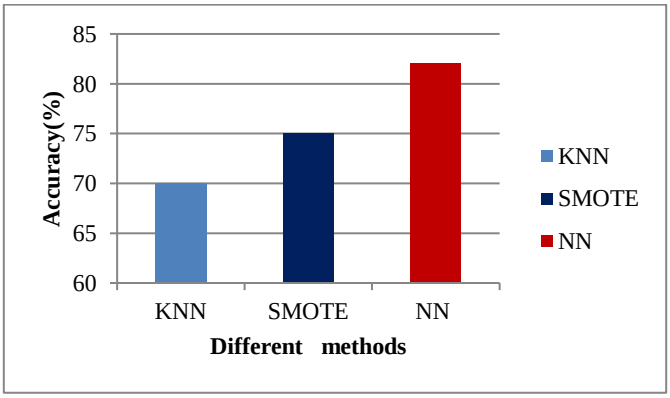

Figure 3: Accuracy Outcomes against Classifiers

Performance measure comparison among suggested NN and current KNN, SMOTE are shown in figure: 3.Interms of accuracy. Through outcomes it is concluded that suggested NN method produces higher accuracy results of 82% while the existing KNN, SMOTE models produces only 70% and 75% accordingly.

VI. Conclusion and Future Work

India is a country where agriculture is extremely important. The nation prospers when the farmers prosper. This survey would assist farmers in sowing the appropriate crop based on environmental and geographical criteria in order to enhance productivity and profit from this method. As a result, farmers may plant the appropriate crop, improving output and enhancing the nation's total production. This article reviews the different crop recommendation model and also reviews about the merits and demerits and found that recent model provides lesser accuracy in decision making. Finally concluded that proper enhancement in deep learning with preprocessing and feature selection models will help to improve the decision making processes which will be used in further research work.

References

[1] S.J. Reashma & A.S. Pillai (2017), "Edaphic Factors and Crop Growth using Machine Learning—A Review", *IEEE International Conference on Intelligent Sustainable Systems (ICISS)*, Pp. 270–274.

[2] D.I. Patrício & R. Rieder (2018). "Computer Vision and Artificial Intelligence in Precision Agriculture for Grain Crops: A Systematic Review", *Computers and Electronics in Agriculture*, Vol. 153, Pp. 69–81.

[3] R. Chetan, D.V. Ashoka & B.A. Prakash (2021), "Smart Agro-Ecological Zoning for Crop Suggestion and Prediction using Machine Learning: An Comprehensive Review", *Advances in Artificial Intelligence and Data Engineering*, Pp.1273–1280.

[4] G. Hegde, V.R. Hulipalled & J.B. Simha (2020), "A Study on Agriculture Commodities Price Prediction and Forecasting", *IEEE International Conference on Smart Technologies in Computing, Electrical and Electronics (ICSTCEE)*, Pp. 316–321).

[5] D.S. Wankhede (2020), "Analysis and Prediction of Soil Nutrients pH, N, P, K for Crop using Machine Learning Classifier: A Review", *International Conference on Mobile Computing and Sustainable Informatics, Springer, Cham*, Pp. 111–121.

[6] N.H. Kulkarni, G.N. Srinivasan, B.M. Sagar & N.K. Cauvery (2018), "Improving Crop Productivity through a Crop Recommendation System using Ensembling Technique", *3rd International Conference on Computational Systems and Information Technology for Sustainable Solutions (CSITSS), IEEE*, Pp. 114–119.

[7] S. Pudumalar, E. Ramanujam, R.H. Rajashree, C. Kavya, T. Kiruthika & J. Nisha (2017), "Crop Recommendation System for Precision Agriculture", *IEEE Eighth International Conference on Advanced Computing (ICoAC)*, Pp. 32–36.

[8] A. Kumar, S. Sarkar & C. Pradhan (2019), "Recommendation System for Crop Identification and Pest Control Technique in Agriculture", *International Conference on Communication and Signal Processing (ICCSP), IEEE*, Pp. 0185–0189.

[9] Y. Cai, K. Guan, J. Peng, S. Wang, C. Seifert, B. Wardlow& Z. Li (2018), "A High-Performance and In-season Classification System of Field-level Crop Types using Time-series Landsat

Data and a Machine Learning Approach", *Remote Sensing of Environment*, Vol. 210, Pp.35–47.

[10] A.T.M. Shakil Ahamed, Navid Tanzeem Mahmood, Nazmul Hossain, Mohammad Tanzir Kabir, Kallal Das, Faridur Rahman & Rashedur M. Rahman (2015), "Applying Data Mining Techniques to Predict Annual Yield of Major Crops and Recommend Planting Different Crops in Different Districts in Bangladesh", *IEEE/ACIS 16th International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD)*, Pp. 1–6.

[11] R.A. Schwalbert, T. Amado, G. Corassa, L.P. Pott, P.V. Prasad & I.A. Ciampitti (2020), "Satellite-based Soybean Yield Forecast: Integrating Machine Learning and Weather Data for Improving Crop Yield Prediction in Southern Brazil", *Agricultural and Forest Meteorology*, Vol. 284, Pp. 107886.

[12] S.K.S. Raja, R. Rishi, E. Sundaresan & V. Srijit (2017), "Demand based Crop Recommender System for Farmers", *IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), IEEE*, Pp. 194–199

[13] S. Sadia, M.B. Propa, K.S. Al Mamun & M.S. Kaiser (2021), "A Fruit Cultivation Recommendation System based on Pearson's Correlation Co-Efficient", *International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)*, Pp. 361–365.

[14] M.S. Suchithra & M.L. Pai (2018), "Improving the Performance of Sigmoid Kernels in Multiclass SVM using Optimization Techniques for Agricultural Fertilizer Recommendation System", *International Conference on Soft Computing Systems*, *Springer*, Pp. 857–868.

[15] A. Goldstein, L. Fink, A. Meitin, S. Bohadana, O. Lutenberg & G. Ravid (2018), "Applying Machine Learning on Sensor Data for Irrigation Recommendations: Revealing the Agronomist's Tacit Knowledge", *Precision Agriculture*, Vol. 19, No. 3, Pp. 421–444.

[16] A. Liu, T. Lu, B. Wang & C. Chen (2020), "Crop Recommendation via Clustering Center Optimized Algorithm for Imbalanced Soil Data", *5th International Conference on Control, Robotics and Cybernetics (CRC), IEEE*, Pp. 31–35).

[17] S.S. Kale & P.S. Patil (2019) "Data Mining Technology with Fuzzy Logic, Neural Networks and Machine Learning for Agriculture", *Data Management, Analytics and Innovation*, *Springer*, Pp. 79–87).

[18] S.B. Kamatchi & R. Parvathi (2019), "Improvement of Crop Production using Recommender System by Weather Forecasts", *Procedia Computer Science*, Vol. 165, 724–732.

[19] F. Löw, U. Michel, S. Dech & C. Conrad (2013), "Impact of Feature Selection on the Accuracy and Spatial Uncertainty of Per-field Crop Classification using Support Vector Machines. *ISPRS Journal of Photogrammetry and Remote Sensing*, Vol. 85, Pp. 102–119.

[20] G. Prabakaran, D. Vaithiyanathan & M. Ganesan (2021), "FPGA based Effective Agriculture Productivity Prediction System using Fuzzy Support Vector Machine", *Mathematics and Computers in Simulation*, Vol. 185, Pp. 1–16.