International Journal of Engineering Science and Advanced Technology (IJESAT)
Vol 24 Issue 2, FEB, 2024

# Future Crop Yield Prediction using Fuzzy C Means and Naive Bayes under Machine Learning

[1 *] Dr.P.Manjula, Associate Professor, Department of Logicnet, Saveetha Institute of Medical And Technical Sciences, manjulasai.sse@saveetha.com

*Corresponding author mail id : manjulasai.sse@saveetha.com

[2] R.Arunkumar, Assistant Professor, Department of CSE, Rajalakshmi Institute of Technology, Chennai, arunkumar.r@ritchennai.edu.in

[3] Dr.V.Anjana Devi , Professor, Department of CSE, Rajalakshmi Institute of Technology, Chennai, anjanadevi.aby06@gmail.com

[4] Dr.N.Kanagavalli, Assistant Professor, Department of CSE, Rajalakshmi Institute of Technology, Chennai, kvalli.818@gmail.com

***Abstract*— Farming is the absolute most significant supporter of the Indian economy. Horticulture crop creation relies upon the season, organic, and monetary reason. The guessing of rural yield is testing and advantageous undertaking for each country. These days, Farmers are battling to deliver the yield on account of flighty climatic changes and definitely lessen in water asset so; we are making a horticulture information. To enhance the existing and as a contribution to our society we are using more productive algorithms to improve the overall efficiency. Hard clustering that is K-means is replaced by soft clustering Fuzzy c means and Naïve bayes is been added to improve the overall performance. Thus the combined methodology will act as a connecting bridge between the farmer and the yield. Results from experiment show that proposed work optimizes the time consumption and increases the overall accuracy.**

***Keywords*— Horticulture, Fuzzy c means, Naïve bayes, CSV**

## I. INTRODUCTION

Currently 11% of the earths land surface is used for agriculture. Prophecy of crop yield has been the major reason behind our economical growth, export income and monetary reason. Prior to the development of technology, prediction and probabilistic calculations was done using the guidance strategy approach. The drawback of the approach is it could not meet the accuracy each and every time. There is a replacement for most of the products we use but food is something which cannot be replaced. In our day to day life scarcity of food we come across might not strike big in our individual life but if it hits hard food scarcity would be a severe human crisis. The major contributor of food is crop and plant. But due to the ongoing changes in climate, atmosphere, water level, crop yield is not at the right pace each year. Here comes the necessity to judge the crop yield beforehand. Whatever the case may be, at the moment with the dramatic development of new computer technologies, various advancement of techniques/algorithms has been developed. This has become a major replacement for old approaches and old techniques . Machine learning plays a major role in the development of technologies since it promotes the ability of a machine to learn on its own and perform the given task .

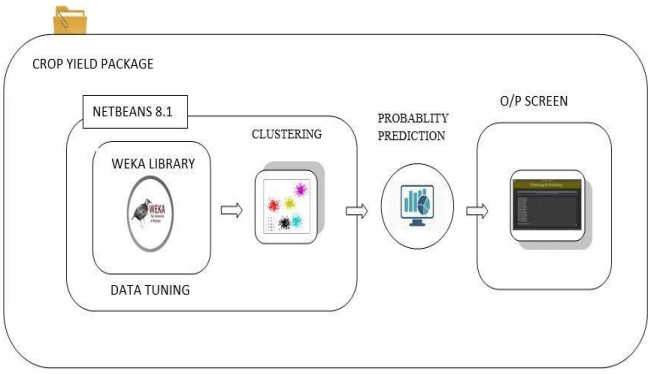


**Figure 1 - Planning Diagram**

Figure 1 explains the flow of the proposed system by combining both the hardware and software specifications and how both the components interact with each other to produce the desired output. It also specifies the library and various tools used in the system.

## II. RELATED WORKS

Main Key Management Challenges in Crop prophecy have been divided into two categories: hard clustering and soft clustering. The prophecy becomes so complex until the clustering is done in proper separation.

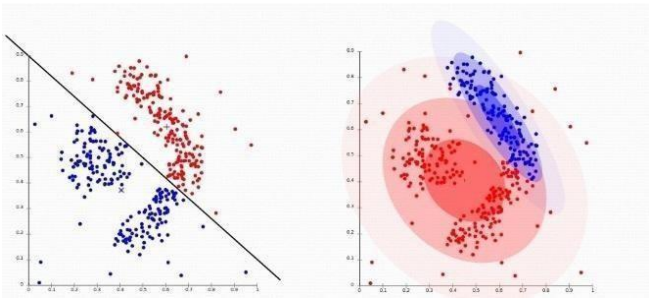

**Figure 2 – Types of Clustering**

Figure 1 demonstrates how the types of clustering are being differentiated and bunched with a detailed graph.

To predict the yield in [2, 4] which uses various ML techniques and SVM, random forest and ID3 are used for investigation. The forecast of farming relies upon boundaries, for example, temperature, soil ripeness, measure of water, water quality and seasons, crop cost.

Vanitha et al. [1] gathers the agricultural data and uses various methodologies such as K-Means, KNN with the tool Jupyter note book and python. It tries various parameters.

Malik et al. [3] explains the investigation on soil properties to foresee richness. The examination has been finished on self - acquired dataset, for three harvests - tomato, potato and bean stew. The harvest yield forecast has been finished utilizing K Nearest Neighbor calculation, Naïve Bayes calculation and Decision Trees classifier.

Shah et al. [5] proposed a recommendation framework, which extract trends by evaluating various parameters such as NDVI Rainfall, Temperature, Air Quality Index (AQI) etc. not only predict the Crop Yield with the help of various Machine Learning and statistical algorithm like XGBoost, Gradient Boosting but also

Sandhya et al. [6] proposed which would assist with foreseeing the gauge of the harvest yield for a particular land in light of the examination of topographical and climatic information utilizing Machine Learning. Relapse models, for example, Decision Tree Regression, K-Nearest Neighbor Regression, Gaussian Process Regression and Support Vector Regression are utilized alongside highlight choice, include scaling, cross approval and hyperparameter tuning procedures to upgrade their exhibition.

Sharma et al. [7] is principally worried about expanding the effectiveness of the current Crop Yield Prediction and diminishing human support to finish framework mechanization. Proposed approach on Crop Yield Prediction involving Hybrid Deep Learning Algorithm for Smart Agriculture.

D. Jayanarayana et al. [8] investigates different ML methods used in the field of harvest yield assessment and gave an itemized examination as far as exactness utilizing the strategies.

Gladence et al. [9] This paper proposes a technique in view of K-Nearest Neighbors (KNN) calculation which identifies the dirt quality and predicts the reasonable yield for development. This paper investigates different ML procedures used in the field of harvest yield assessment and gave a nitty gritty examination as far as exactness utilizing the strategies.

Bhanumathi et al. [10] This paper comes with a system model to be precise and accurate in predicting crop yield and deliver the end user with proper recommendations about required fertilizer ratio based on atmospheric and soil parameters of the land which enhance to increase the crop yield and increase farmer revenue.

Reshma et al. [11] proposed IoT framework is made out of pH sensors, Humidity and temperature sensors, Soil dampness sensors and so forth. For the suggesting framework, the SVM and Decision Tree calculation is proposed to get the harvest appropriate for the given soil information and assists with improving the development utilizing an upgraded cultivating process.

The main aim of these papers [13, 14] is to predict crop yield using area, yield, production, area and weather. These paper explores various ML techniques such as Decision Tree, Linear Regression, Lasso regression, and Ridge Regression have been applied to estimate the crop yield.

Kale et al. [15] This paper considered for study are development region, crop, state, locale, season, year and creation or yield for the time of 1998 to 2014. This examination portrays the improvement of an alternate harvest yield forecast model with ANN, with 3 Layer Neural Network. The regressive and forward engendering strategies are utilized.

## III. PROPOSED SYSTEM

Our proposed system is based on the combination of Fuzzy C Means, which is a type of clustering which specifically denotes technique known as Soft Clustering and Naïve bayes which is used to deliver the probability based on various criteria. Soft Clustering is an upgradation with the existing systems and in addition Naïve bayes join hands to improve the overall accuracy and efficiency of the system. Fuzzy C Means Clustering belongs to soft clustering in which a instance may belong to one or more cluster so that the output will be efficient. For example there is a group of 100 persons and persons who weigh more than 80 would be overweight and less than 60 would be less weight. When soft clustering is performed persons who weigh between 60-80 would be clustered into a separate instance so that no value gets neglected to maintain the at most accuracy. Naive bayes methodology is performed on the clustered instance to find the best crop that can be yield and the value of production in tonnes will also be identified. The proposed system uses Netbeans as an IDE and java as the programming language. To make it easy user friendly interface is built and swing buttons are added to make it more attractive and easily accessible. When the proposed system is being executed button stating load Tamil Nadu agriculture yield date has to be clicked and the dataset gets pre-process, feature extraction takes place. Clustering and prediction button has to be clicked so that the model performs it task and the yield value will be given in tonnes and accuracy value will also be visible as the output so that the farmer makes best choice. Our proposed system is done in a way to enhance the existing system and to produce best accuracy, efficiency and less time complexity.

### III.1 Data Pre-processing

Data pre-processing is described as processing of the unprocessed raw data into suitable form so that it would be exactly suitable to process in the desired model. In the stage of pre-processing also unnecessary piece of information would be removed so that the time consumed and the overall time complexity will also become less. Raw material in the harvest information which is present in the form of CSV type is converted into whole numbers so that it would be suitable for the proposed model and this process denotes the pre-processing stage. For example in the proposed system data such as the district names and the crop names that are present in the CSV type is converted into numeric. The pre-processed data will be appropriate to be clustered. Pre-processing is a vital procedure to obtain the most accurate and efficient result.

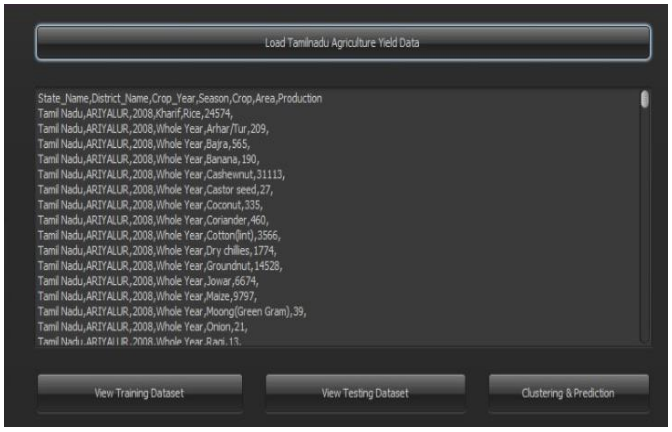


**Figure 3 - Data Pre-Processing**

Figure 2 portraits by clicking on the Load Tamil Nadu agriculture Yield Data Button the dataset collected is loaded in CSV type into the IDE to get executed.

### III. 2 Feature Selection

Feature selection plays a major contribution in the overall efficiency of the proposed system as well as any other system that is going to produce as a output/resultant. Crop prophecy may contain various input attributes such as crop, year, rainfall, temperature, soil, seed strength and lot more. But the main idea behind the proposed system is to pick the least number of input attributes so that the processing time and time complexity would be less at the same time the attributes will be the top vital attributes and produce the best accuracy. Misleading of selection of wrong attributes may lead to results with less accuracy and might affect the system. Keeping these scenarios in mind, In the proposed system crop, year, season, production in tonnes, area is taken as vital parameters to obtain maximum accuracy and efficiency.

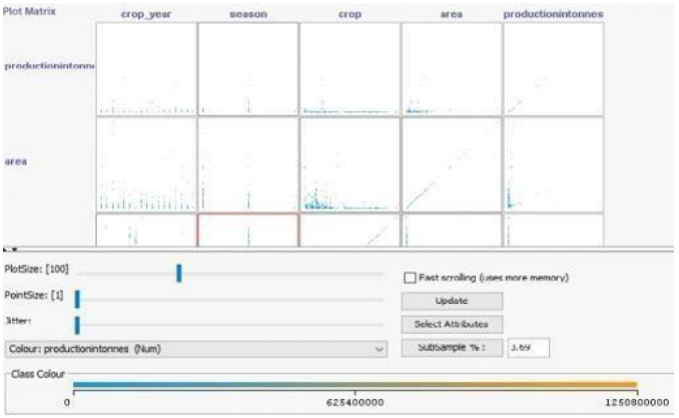


**Figure 4 - Feature Selection**

Figure 3 lists out various attributes like crop year, production in tonnes, crop, area, season that can be selected and also as features to increase the plot size.

### III.3 Feature Extraction

Feature Extraction is the process of converting the raw data into numeric values. For example in the proposed system the CSV type of raw data is converted into whole numbers in a way it is suitable for the proposed model to process which is known as feature extraction. The extracted features is present in the form of dataset which will be trained and tested so that the process will be able to reach at most accuracy.

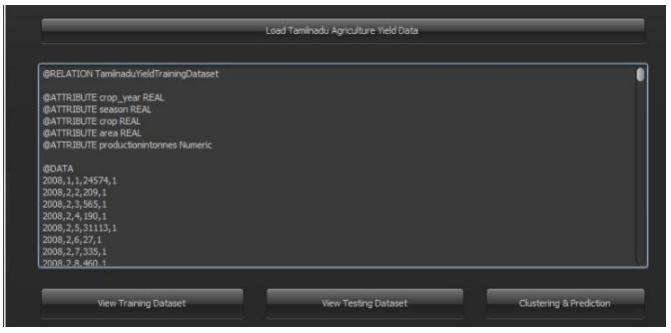


**Figure 5 - Feature Extraction**

Figure 4 displays the extracted features from CSV type (given in figure 2) so that the proposed model can process on it.

## IV. CLASSIFICATION ALGORITHM

### IV.1 Fuzzy c means

Clustering is a process of separating and grouping the data points or objects with some given criteria. There are two major types of clustering which is Hard Clustering and Soft Clustering. Fuzzy C Means comes under soft clustering technique which is used in the proposed system to overcome the disadvantages of Hard Clustering and also combined with naive bayes to produce better results.

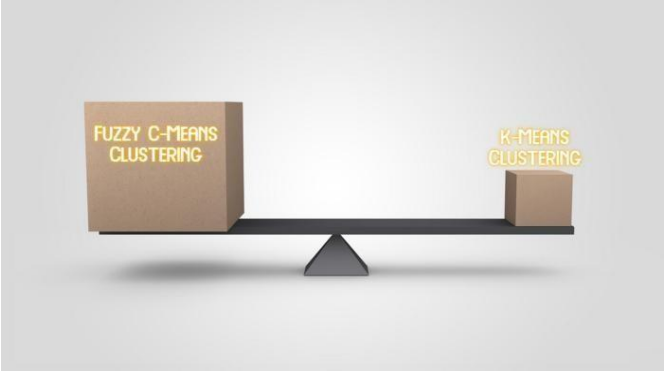


**Figure 6 - Comparison of Clustering**

Figure 5 portraits that Soft Clustering has more advantages than Hard Clustering being the reason to use in our system.

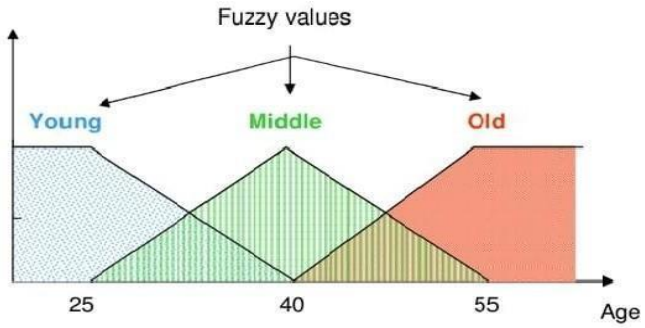


**Figure 7 - Fuzzy C Means Clustering**

Figure 6 explains how the data points are clustered without neglecting of any data points in Fuzzy C Means algorithm.

For example there is a group of 50 persons and persons whose age is less than 25 would be young and more than 55 would be less old. When hard clustering is performed the values between the range 25 and 55 would be neglected or it would be rounded to the nearby wholesome value. This might not be a big deal when we are dealing with small amount of data but when it comes to large amount of data it will affect the accuracy and lead to inappropriate results. On the other hand soft clustering is performed in which persons whose age is between 25-55 would be clustered into a separate instance as medium as described in figure 6, so that no value gets neglected to maintain the at most accuracy.

After the conversion of CSV data type into numeric values the testing data set is ready to be clustered. The dataset clustered us made into many instances on basis of crop name and the tons yielded. The testing dataset consist of numeric value where the district name, crop name, area is converted into numeric value. Production yield in tonnes and year is already in numeric format. For example the crop rice would be clustered with respective years that are present in the testing data set and hence all the crops will be clustered. They are separated individually hence the future yield could be predicted. Thus the clustering is done and naive bayes is ready to get executed on the resultant data set to predict future yield in tons.

**Fuzzy C Clustering Formula:**

$$J_f(C,m) = \sum_{i=1}^{c} \sum_{k=1}^{N} (u_{k,i})^m \, d_{k,i}$$

| | |
|---|---|
| C | Number of clusters |
| N | Number of data points |
| $U_{i,k}$ | Fuzzy membership of the k-th point to the i-th cluster |
| $d_{k,j}$ | Euclidean distance between the data point and the cluster center |
| $m \in (1,\infty)$ | Fuzzy weighting factor which defines the degree of fuzziness of the results. |

### VI.2. Naïve bayes

Naive bayes is a methodology or an algorithm based on probability. After the dataset is being clustered using Fuzzy C Means the next vital step is to decide which crop has to be yielded. Hence Naive Bayes technique is used to determine which crop yielded and the outcome is displayed in tonnes. The probabilistic operation would be performed on the clustered dataset. For example the instance of crop rice is present and naive bayes is performed on the production in tonnes for the previous years. From there by speculating that the buildings should be automatically distributed, Naive Bayesian conducted his experiments using the calculations used to obtain significant results in the dumping.

**Naïve bayes Formula:**

$$P(c|x) = P(x|c)\ P(c) / P(x)$$

| | |
|---|---|
| P(c\|x) | Posterior probability of class(c,target) given predictor |
| P(c) | Prior probability of class |
| P(x\|c) | Likelihood which is probability of predictor given class |
| P(x) | Prior probability of predictor |

The central improvement of the Naive Bayes investigation is it tends to be faster to prepare as well as order. Additionally Naive Bayes isn't delicate to insignificant highlights. Its premise is genuine and discrete information and furthermore oversees streaming information too. The Credulous Bayes take advantage of the opportunity for the harvest to be completed in those conditions. So in the corresponding results the chances are determined and the most likely yield is selected for the final harvest. In exactness comparison the naive bayes has the biggest number of rate when contrasted and fuzzy c means.

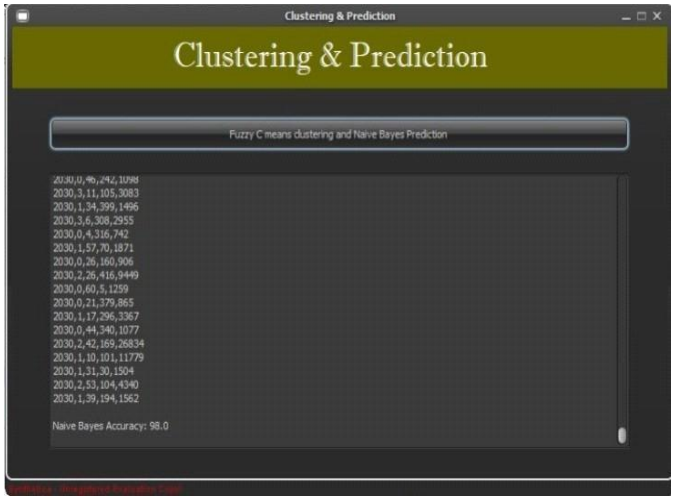


**Figure 8 - Clustering and Prediction**

Figure 7 shows the output after both clustering and prediction is being predicted. It also displays the overall accuracy and production in tonnes for the upcoming years so that the farmer can choose the crop wisely to attain at most profit and yield.

## V. EXPERIMENTAL ANALYSIS

Real CSV generic information that includes a record for each line of text that matches the given long-term information. Information for creating a CSV vintage creation with a record for each line of text contains information for a specific month of the year. In the pre-processing phase, however there are many limited limits accessible to the polluted nature database, the most important non-essential points that need to be addressed in a review are ignored and important factors are considered. Consequently the two distinct sorts of chronicled records are pre-processed as well as joined together so it tends to be utilized for this review. The time series recorded information more than 10 decade years is taken into study for this analysis. Preparation and testing, database classification was performed where 60% of the database was used for preparation and 40% of the database was used for testing. By utilizing grouping evaluation, preparation dataset is utilized and creation of models done. The generated model is used in the test database to check the accuracy. Thus the two methodologies that has been used in the proposed system paves way to get the most accurate result.

## VI. CONCLUSION

Fuzzy C Means and Nave Bayes are the two methodologies used in the proposed system from which the investigation of time series for the crop yield has been revealed. From the proposed method by examining the result it is found that it meets the accuracy as excepted. Further research of this review can lead to the improvement of the efficiency of the result by alternating the existing method. Horticulture as well as the economy can be developed only if the yield prediction id done to at most accuracy. To be more productive constant examination has to be and better techniques or methodologies has to be discovered to improve the efficiency or accuracy even more and to work on new ideas. The current undertaking is the beginning for an supplemental investigation in gauging.

## VII. FUTURE ENHANCEMENT

The acquired outcome would be useful to the ranchers to understand the yield that would be happening in the future. So that the ranchers can predict and select better crop to harvest that will give high return and furthermore tell them the effective utilization of manure so the rancher can involve just the expected measure of composts for the field. The proposed system might also be developed into an application to make sure it reaches each type of audience. The application would have a drop down menu to select number of years and it would ask the end user to enter the crop and respective yield that took place and can deliver the predicted yield for the upcoming years as an output to the end user. This way our project can assist the ranchers with developing the harvest thatwill gives them better yield and harvesting.

## REFERENCES

[1] Vanitha, C. N., N. Archana, and R. Sowmiya. "Agriculture analysis using data mining and machine learning techniques." 2019 5th international conference on advanced computing & communication systems (ICACCS). IEEE, 2019.

[2] Nigam A, Garg S, Agrawal A, Agrawal P. Crop yield prediction using machine learning algorithms. In2019 Fifth International Conference on Image Information Processing (ICIIP) 2019 Nov 15 (pp. 125-130). IEEE.

[3] Malik, Pranay, Sushmita Sengupta, and Jitendra Singh Jadon. "Comparative analysis of soil properties to predict fertility and crop yield using machine learning algorithms." 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence). IEEE, 2021.

[4] Sajja, Guna Sekhar, Subhesh Saurabh Jha, Hicham Mhamdi, Mohd Naved, Samrat Ray, and Khongdet Phasinam. "An Investigation on Crop Yield Prediction Using Machine Learning." In 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 916-921. IEEE, 2021.

[5] Shah, Avnika, Rhea Agarwal, and B. Baranidharan. "Crop Yield Prediction Using Remote Sensing and Meteorological Data." In 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 952-960. IEEE, 2021.

[6] Sandhya, V., & Padyana, A. (2021, November). Machine Learning based Crop Yield Prediction on Geographical and Climatic Data. In 2021 Sixth International Conference on Image Information Processing (ICIIP) (Vol. 6, pp. 186-191). IEEE.

[7] Sharma, Avdesh Kumar, and Anand Singh Rajawat. "Crop Yield Prediction using Hybrid Deep Learning Algorithm for Smart Agriculture." 2022 Second International Conference on Artificial Intelligence and Smart Energy (ICAIS). IEEE, 2022.

[8] Reddy, D. Jayanarayana, and M. Rudra Kumar. "Crop yield prediction using machine learning algorithm." 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE, 2021.

[9] Gladence, L. M., Reddy, K. R., Reddy, M. P., & Selvan, M. P. (2021, June). A Prediction of Crop Yield using Machine Learning Algorithm. In 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI) (pp. 1072-1077). IEEE.

[10] Bhanumathi, S., M. Vineeth, and N. Rohit. "Crop yield prediction and efficient use of fertilizers." 2019 International Conference on Communication andSignal Processing (ICCSP). IEEE, 2019.

[11] Bang, Shivam, Rajat Bishnoi, Ankit Singh Chauhan, Akshay Kumar Dixit, and Indu Chawla. "Fuzzy logic based crop yield prediction using temperature and rainfall parameters predicted through ARMA, SARIMA, and ARMAX models." In 2019 Twelfth International Conference on Contemporary Computing (IC3), pp. 1-6. IEEE, 2019.

[12] Kalimuthu, M., P. Vaishnavi, and M. Kishore. "Crop prediction using machine learning." In 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT), pp. 926-932. IEEE, 2020.

[13] Kavita, Ms, and Pratistha Mathur. "Crop yield estimation in India using machine learning." In 2020 IEEE 5th International Conference on Computing Communication and Automation (ICCCA), pp. 220-224. IEEE, 2020.

[14] Kale, Shivani S., and Preeti S. Patil. "A machine learning approach to predict crop yield and success rate." *2019 IEEE Pune Section International Conference (PuneCon)*. IEEE, 2019.

[15] Kumar, Y. Jeevan Nagendra, V. Spandana, V. S. Vaishnavi, K. Neha, and V. G. R. R. Devi. "Supervised Machine learning Approach for Crop Yield Prediction in Agriculture Sector." In 2020 5th International Conference on Communication and Electronics Systems (ICCES), pp. 736-741. IEEE, 2020.

[16] Reddy, Kasara Sai Pratyush, Y. Mohana Roopa, Kovvada Rajeev LN, and Narra Sai Nandan. "IoT based smart agriculture using machine learning." In 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA), pp. 130-134. IEEE, 2020.

[17] Keerthana, Mummaleti, K. J. M. Meghana, Siginamsetty Pravallika, and Modepalli Kavitha. "An ensemble algorithm for crop yield prediction." In 2021 Third International Conference on Intelligent

Communication Technologies and Virtual Mobile Networks (ICICV), pp. 963-970. IEEE, 2021.

[18] Medar, Ramesh, Vijay S. Rajpurohit, and Shweta Shweta. "Crop yield prediction using machine learning techniques." In 2019 IEEE 5th International Conference for Convergence in Technology (I2CT), pp. 1-5. IEEE, 2019.

[19] Kanagavalli, N., Baghavathi Priya, S., Jeyakumar, D, "Design of Hyperparameter Tuned Deep Learning based Automated Fake News Detection in Social Networking Data",Proceedings - 6th International Conference on Computing Methodologies and Communication, ICCMC 2022, 2022, pp. 958–963

[20] Sandhya, V., and Ajith Padyana. "Machine Learning based Crop Yield Prediction on Geographical and Climatic Data." In 2021 Sixth International Conference on Image Information Processing (ICIIP), vol. 6, pp. 186-191. IEEE, 2021.