ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

ADVANCED DATA ANALYSIS AND CLUSTERING METHODS FOR CROP YIELD PREDICTION USING ARTIFICIAL SYSTEM

[1]Dhara Jaya Soniya [2]Dr. Prasuna Grandhi [3]Dr. A. Tirupataiah
[1]M. Tech Scholar, Dept. of CSE, St. Ann's College of Engineering & Technology, Chirala.
[2]Associate Professor, Dept. of CSE, St. Ann's College of Engineering & Technology, Chirala.
[3]Associate Professor, Dept. of CSE, St. Ann's College of Engineering & Technology, Chirala.
e-mail: grandhiprasuna@gmail.com

ABSTRACT: The science and skill of cultivating plants and fauna are referred to as agriculture. 60.45% of Indian land is used for farming, which gives the country the second-highest global ranking. These agricultural economy-related problems result in higher crop yield. Crop Yield Prediction is crucial in today's agriculture market (CYP), which is expanding quickly. Selected features and machine learning techniques are necessary for accurate prediction. An electronic agricultural record (EAR), which is effectively recommended by agronomists as a vital component of a smart crop system, is intended to integrate many different datasets. Dirichlet Allocation and Artificial Neural Network classification algorithms are used to determine the causes of a given plant disease. To predict a suitable yield, the climate and soil conditions are taken into account.While LSTM and RNN are employed as Deep Learning algorithms, the SVM is implemented as a Machine Learning method. There are numerous clustering techniques, including k-Means, Expectation-Maximization, Hierarchical Micro Clustering, Density-Based Clustering, and Weight-based Clustering, which are briefly discussed. A novel clustering method, Epsilon Density-Based Prediction (EDBP), is also suggested to update the best crop production prediction. In order to anticipate yield, this paper uses ANN with cascade-forward back propagation and Elman back propagation. We employed Regression methods, Decision trees, Naive Bayes, SVM, K-Means, Expectation-Maximization (EM), and AI approaches (LSTM, RNN) together with machine learning and deep learning algorithms. For predicting agricultural yield the Random Forest algorithm has a training accuracy of 99.27%. With 95% accuracy and 92% sensitivity, the suggested system produces good results.

Index Terms: Agriculture, Crop Prediction, Machine Learning, Deep Learning, SVM, LSTM, RNN, Suitability Assessment, Temporary Crops, K Means, Machine Learning, Prediction**INTRODUCTION**

The population has been growing exponentially over the past few years. Parallel to the rapidly expanding population, there has been a massive growth in associated needs [1]. Resources for crop production, such as available freshwater and cropland, are severely constrained, which alters the issue [2]. Agriculture used to be categorised by increased production, the replacement of man-made fertilisers and insecticides, and the allowance of additional land [3].One of the most widely grown and farmed crops worldwide, including in China, India, the United States, Brazil, and Russia, is tobacco. The suggested structure calculates for creating predictions similar to multiple linear regression in order to recognise information using artificial intelligence (AI). Additionally, it is treated under the conditions mentioned, including [4].Anyone is surprised by the technology's

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

rapid advancement in agriculture. Five farmers advocate using technology to boost productivity and control costs [5]. By taking into account various rice characteristic, the Epsilon Density Based Prediction (EDBP) clustering algorithm is created for crop yield prediction and to maximise production. Production agriculture in the agricultural sector is dependent on numerous biological, climatic, economic, and human elements, all of which interact intricately [6]. Decision Tree (DT), Support Vector Machine (SVM), Linear Regression (LR), Random Forest (RF), Support Vector Regression (SVR), Artificial Neural Network (ANN), Deep Learning (DL), Convolution Neural Network (CNN), Bayesian Network (BN), K-Means clustering, and K-Nearest Neighbour (KNN) are some of the most popular new machine learning techniques [7]. The TSVM classification approach predicts the plant diseases. The probability values between soil pictures and diseased plant images are used in ANN classification using the plant illnesses and expected and soil image attributes [8]. proposed cluster segmentation based on support vector machines (SVM) and machine learning. This system deals with plant orientation and lighting models. Deep learning-based advanced model for crop protection plant detection and counting [9].

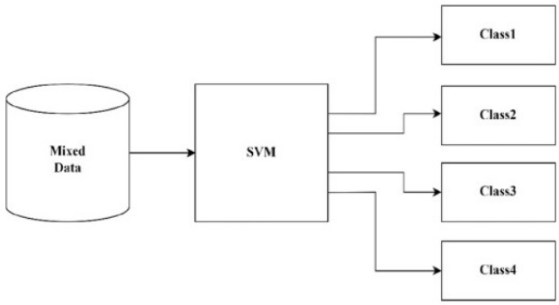


Figure 1. Presents the Graphical Presentation Model

1. RELATED WORK

To reduce the number of agents and intermediary hops between clients and ranchers, which further supports the rancher, IOT-based models are deployed. The research project ultimately takes its cues from the publication [10]. One of these studies showed how to forecast rice crop yield using the Support Vector Machine (SVM), a machine learning method. The objective is being reached by using the precipitation (mm), minimum, average, and maximum temperatures as climatic parameters. The choice and optimisation of plantation dates are included in the fuzzy query model for crops [11]. The expectation-maximization technique, the density-based approach, the weight-based method, and the hierarchical clustering algorithm are all clustering methods.In order to uncover critical information in agricultural data sets, clustering algorithms play a vital role in data retrieval and Text Mining Artificial Neural Networks, K Means, and K Nearest Neighbours algorithms [12]. The best approaches now in use are semi-automated ones, as completely automated methodologies are being developed. A well-established subclass of semi-automated models, machine learning has proven benefits in a variety of application fields [13]. It changes academic industries to develop an effective and powerful machine learning approach outputs is generalised in the best possible answer. The Heterogeneous Ensemble Learning Environment (HELE) is created with clarity and great robustness [14].The system offered effective management features for determining the right amount of fertiliser, and a water

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

system was developed for updating crop plans by utilising data on cropping patterns, rainfall, water status, and land usage [15].

2. SYSTEM MODEL

The influence of climate change on global crop output is likely due to the rapid rise of the world's population, which has an impact on crop yield estimation and crop monitoring [16]. The EM algorithm is an iterative refinement method used to identify the probability distribution's parameters. In accordance with the Expectation and Maximization process, the parameters are refined iteratively. Real-world datasets that are susceptible to noise are the best for addressing [17]. Future farmer data and crop output will make it easier to identify and prevent crop losses. Crop yield prediction is a significant issue in the agriculture sector [18]. Farmers consistently obtain yields that are below their expectations. When building expert systems to find new agricultural knowledge and increase farm productivity, smart farming uses statistical and data mining algorithms on historical data. It also starts with tools for farmers. By 2023, the market for agricultural analytics will have grown by more than 110%, from $580 million in 2018 to $1.236 million [19].

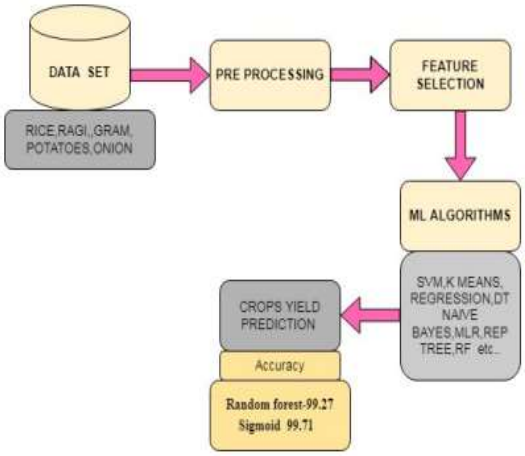


Figure 2. Predicting Crop Yield Prediction.

3. PROPOSED SYSTEM

The proposed approach employs machine learning and deep learning techniques to run an experiment on a crop dataset in order to determine the best crop production strategy [20]. With this strategy, harvests can be precisely forecast. SVM implementation is under machine learning, but LSTM and RNN implementation falls under deep learning [21]. A machine learning model is used to estimate agricultural yield based on climate variables. Today's research has improved the user web page for a software application called Crop Advisor, which forecasts the impact of climatic conditions on crop yields.The framework of the new model is provided with accurate and correct outcomes by AI computations like Random Algorithm (RFA) and Back Propagation Algorithm [22].

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

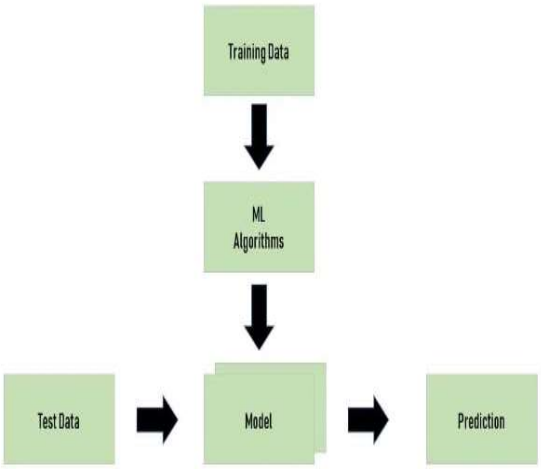


Figure3. Proposed Model used to Predict Crops

4. RESEARCH METHODOLOGY

The Epsilon Density Based Prediction (EDBP) clustering method is used for finding the crop yield analysis. The implementation procedure

Let X = (x1,x2,x3,....., xn) be the data set with number of attributes such as, Area of Sowing, Cloud Cover, Pressure [23],

Step 1: Divide the data set into 70% training set and 30% test set using leave one out method without selecting dependent variable.

Step 2: Select each record of test set xi to find the distance from each record of training using Euclidean distance.

Step 3: Fix a value of epsilon = 2,3,4,5, ..., n where n >0 and n

Step 4: In step 2, We get a matrix for all records of the test called dataset.

Step 5: From these arranged records in Epsilon Density, €=2,3,4.The dependent variable values of these Epsilon Density, € are selected as a set 'M' for each record in the test set.

Step 6: For all epsilon values is measured.

Step 7: The model is best for prediction for the find values in arrange dataset.

A. SUPPORT VECTOR MACHINE (SVM)

Step 1: Import the required dataset.

Step 2: Load input dataset.

Step 3: Choose the required many features in the dataset.

Step 4: Total SVM values with the help of original data.

Step 5: Take value for the regularization parameter.

Step 6: Finally data object of SVM classifier is generated [24].

B. LONG-SHORT TERM MEMORY (LSTM)

Step 1: Take new neural network in sequence of layers.

Step 2: Total network requiresand specified function.

Step 3: The network requires the specified training data total input patterns and matching output

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

patterns array.
Step 4: Find network in training data. The model is validation dataset [25].
Step 5: To required predictions in format given by the network output layer.

C. RECURRENT NEURAL NETWORK (RNN)

Step 1: The network provided with single time step of input.
Step 2: With the help of the back state and the current input take current state.
Step 3: The current state turns out of the next time step.
Step 4: New number of time is depending on information is joined in entire back states.
Step 5: The completion of the entire time steps in final state is used computes the output.
Step 6: In order toupdating weights the error is back-propagated towards the network is RNN is trainee [26]

D. ANN WITH BACK PROPAGATION

Machine learning is a key technology that performs non-linear operations recursively for many applications. The most crucial step in this process is to properly train a neural network. "Back-propagation" is vital to this training but is very ambiguous to most beginners to deep learning [27].

Back-propagation is a neural network training technique where the weights of a neural network are fine-tuned and then retrieved from the previous iteration of an epoch. The technique is trustworthy in ensuring adequate weight tuning, which changes decrease error rates and increases generalisation.

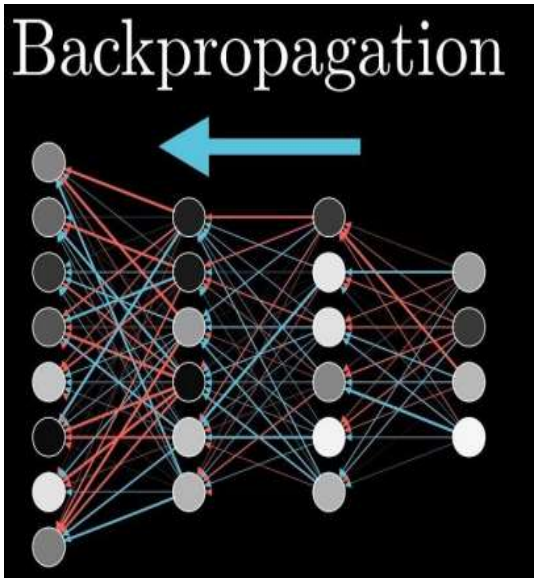


Figure:4. ANN with Back propagation

A model is established under training that performs the functionality of XOR using three hidden units and two inputs
f(a) = A
A is hypothesis function to find input to the activation function the function is typical and popular one: h(X) = W0.X0 + W1.X1 + W2.X2 (W, X) is function relevant to the logistic regression cost function and it looks complex a little bit but is simple actually

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

E. CLUSTER BASED SEGMENTATION

A binary segmented region's boundary is found using an edge detection technique to increase the pace at which tobacco plants may be detected. The centroid of each region is utilised as a feature point for clustering to discover the centre of each plan as the major goal of this method in the plant margin angle of all bright regions inside the plants finding modify brighter regions is taken from plant regions is specified template.

F. IMAGE PRE-PROCESSING

The input images of different plants and soil collected by using the digital camera with a desired resolution for better quality The captured images is transmitted in wired or wireless network to the image processing unit in further processing the collected images is pre-processed for removing the noise in different disturbances from the images and update the features for prediction method.

The training algorithm is containing the following stepsheadings.

1. Set parameter C and $C*$
2. Used inductive SVM in training data set to change initial classifier
3. Set the number of positive dataset based on rule
4. Compute search function values total dataset
5. Label dataset highest decision function is positive
6. Set temporary effect factor $Ctmp*$
7. Retrain the support vector machine over total dataset
8. Switch labels of every labeled unlabeled is certain rule to make the value of objective function by using (5)
9. Repeat the process until dataset is satisfying the switching condition
10. Increase the value of $Ctmp*$ then move to step 7
11. ($Ctmp*>C*$)
12. Stop
13. Optimize the output

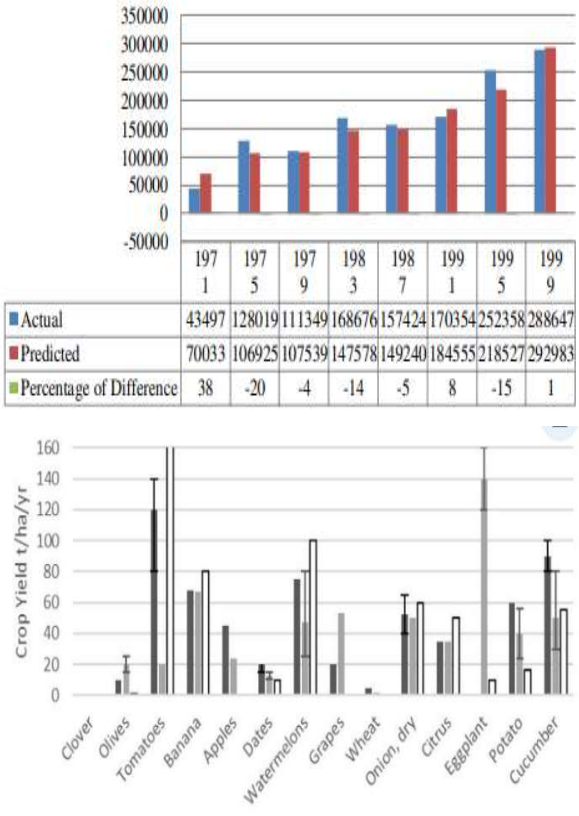
The data is compared with training data type of plant disease is predicted effectively to the required water content level for the specific type of plant is predicted based on the features extracted from the soil images by using LDA technique.

5. RESULTS ANALYSIS

The dataset of gathered crops is uploaded for the research's implementation. Importing the required libraries and packages comes first, and then data searching comes next. The data is divided into test data and training data. Finally, a model is built in which the best crop is produced on a specific soil and necessary AI algorithms are utilised in return. In terms of precision and accuracy, the performance of both plant disease prediction and the aetiology of those illnesses is assessed for both existing and planned methodologies. The outcomes of the prediction of plant disease causes are contrasted with those of the LDA-based technique and similarity measure-based technique. It is necessary to increase the size of the training sample, tweak the hyper parameter, simplify the model, and disable system regularization in order to reduce overfitting.

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

In our study, a data augmentation strategy has been employed to expand and give ML algorithms with an acceptable number of training examples. Even if the dataset we use has 500 photos, it is still extremely little. We use data transformation and intensity transformation, which are discussed in the preceding section, to improve prediction performance. The dataset size is reduced after data augmentation and resizing.Gather information about what kind of crops majorly grown in India. Arrange the data in the form of bar graph



	1971	1975	1979	1983	1987	1991	1995	1999
Actual	43497	128019	111349	168676	157424	170354	252358	288647
Predicted	70033	106925	107539	147578	149240	184555	218527	292983
Percentage of Difference	38	-20	-4	-14	-5	8	-15	1

Figure 5. Comparison between Actual production and Predicted values

The first bar the red bar the accuracy as 93% when ANN and Random Forest algorithms is used. Whereas the second bar the yellow the accuracy as 97% when applying LSTM, RNN and SVM algorithms together the second result gives better accuracy.

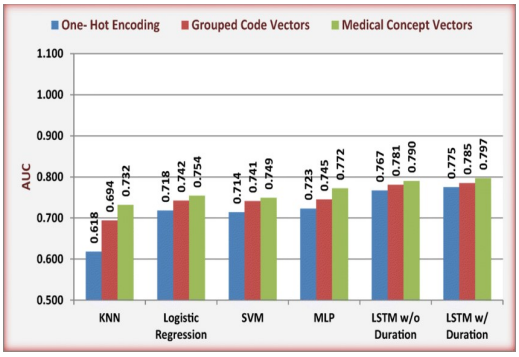


Figure 6. Performance Graph displaying accuracy level

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

6. CONCLUSION AND FUTURE WORK

The proposed model is built using AI algorithms to address the issue of farmers losing money on their farms as a result of their ignorance cultivate in various soil and weather conditions. Machine learning (SVM) and deep learning (LSTM, RNN) methods are used to build this model. Utilizing previously obtained data, deep neural networks have been used to produce successful results in terms of crop productivity. Additionally, they are able to anticipate crop yields accurately planted hybrids in unfamiliar places with established weather patterns.Random forest has 99.27% of the machine learning and deep learning approaches at the training set and 98.25% at the testing set. The EAR is changed and scaled to accommodate new datasets and various. Our suggested monitoring system is used to enhance the crop production monitoring system in terms of effective disease control and irrigation. In the future, the forecast method for tracking the plant's growth level will be taken into account to increase agricultural productivity.To determine the relationship between the components of fertilizers, soil adjuvants, and water requirements on boosting crop output. Future research will focus on developing advanced models that should be more precise and well-described in order to overcome this restriction.

7. REFERENCES

[1] R. Jahan, ‘‘Applying naive Bayes classification technique for classification of improved agricultural land soils,’’ Int. J. Res. Appl. Sci. Eng. Technol., vol. 6, no. 5, pp. 189–193, May 2018.

[2] Suma N, Samson S R, Saranya S, Shanmugapriya G and Subhashri R February 2017 IOT Based Smart Agriculture Monitoring System IJRITCC

[3] Dhivya B, Manjula, Bharathi S and Madhumathi March 2017 A Survey on Crop Yield Prediction based on Agricultural

[4] Ravichandran G, Koteeshwari R S 2016 Agricultural Crop Predictor and Advisor using ANN for Smartphones IEEE

[5] Dr. Ratna Raju Mukiri2018 Estimate Requirement of Online Report Analysis using Random Forest Regression

[6] A.Awan and M.Md.Sap," A framework for predicting oil-palm yield from climate data," International Journal of Information and Mathematical Sciences,2012.

[7] S. Kim and Wilbur, " An EM clustering algorithm which produces a dual representation," In: IEEE 10th International

[8] E. Manjula*, S.Djodiltachoumy “Analysis of Data Mining Techniques for Agriculture Data” International Journal of Computer Science and Engineering

[9] D Ramesh, B Vishnu Vardhan “Crop Yield Prediction Using Weight Based Clustering Technique’ International Journal of Computer Engineering

[10] B. Vishnu Vardhan, D. Ramesh and O. SubhashChanderGoud, “Density Based Clustering Technique on Crop Yield Prediction”

[11]. G. M. Fuady, A. H. Turoobi, M. N. Majdi, M. Syaiin, R. Y. Adhitya, I. Rachman, R. T. Soelistijono. Extreme learning machine and back propagation neural network Devices (ISESD) 2017, 46-50.

[12]. F. Balducci, D. Fomarelli, D. Impedovo, A. Longo, G. Pirlo. Smart Farms for a Sustainable

ISSN:2731-538X | E-ISSN:2731-5398
Vol. 18 No. 01 (2024)

Annual Conference, Bari 2018, 1-6.
[13]. S. Gertphol, P. Chulaka, T. Changmai. Predictive models for Lettuce quality from the Internet of Things-based hydroponic farm. 2018
[14]. Z. Doshi, S. Nadkarni, R. Agrawal, N. Shah. AgroConsultant: Intelligent Crop Recommendation System Using Machine Learning Algorithms. , India 2018, 1-6.
[15]. B. Fabrizio, I. Donato, P. Giuseppe. Machine Learning Applications on Agricultural Datasets for Smart Farm Enhancement. Machines 2018, 6, 38
[16] Bolten, J. D., Crow, W. T., Zhan, X., Jackson, T. J., and Reynolds, C. A. "Evaluating the utility of remotely sensed soil moisture retrievals for operational
[17] Barbedo, J. G. A. "A novel algorithm for semiautomatic segmentation of plant leaf disease symptoms using digital image processing,", 41(4), pp. 210-224, 2016.
[18] Han, L., Haleem, M. S., and Taylor, M. "A novel computer vision-based approach to automatic detection and severity assessment of crop diseases,"
[19] Xiaoxia, Y., and Chengming, Z. "A soil moisture prediction algorithm base on improved BP," In Fifth International Conference on Agro-Geoinformatics
[20] Sung, W. T., Chung, H. Y., and Chang, K. Y. "Agricultural monitoring system based on ant colony algorithm with centre data aggregation"
[21] Teimouri, N., Omid, M., Mollazade, K., and Rajabipour, A. "A novel artificial neural networks assisted segmentation algorithm for discriminating
[22] Z. Fan, J. Lu, M. Gong, H. Xie, and E. D. Goodman, "Automatic tobacco plant detection in uav images via deep neural networks," IEEE Journal of Selected Topics [23] Y. Chen et al., "Citrus tree segmentation from UAV images based on monocular machine vision in a natural orchard environment,"
[24] B. Neupane, T. Horanont, and N. D. Hung, "Deep learning based banana plant detection and counting . 2019, doi: 10.1371/journal.pone.0223906.
[25] X. Sun, J. Peng, Y. Shen, and H. Kang, "Tobacco plant detection in RGB aerial images," Agriculture,.
[26] K. Aitelkadi, H. Outmghoust, S. Laarab, K. Moumayiz, and I. Sebari, "Detection and counting of fruit trees from RGB UAV images by convolutional neural networks.
[27] F. Gao et al., "Multi-class fruit-on-plant detection for apple in SNAP system using faster R-CNN,".
[28] S. K. Behera, A. K. Rath, and P. K. Sethy, "Fruits yield estimation using Faster R-CNN with MIoU," Multimedia Tools and Applications, vol. 80, no. 12, pp. 19043–.
[29] J. Liu and X. Wang, "Tomato diseases and pests detection," Frontiers in Plant Science, vol. 11, pp. 1–12, Jun. 2020, doi: 10.3389/fpls.2020.00898.