A Comprehensive Survey on Estimating Crop Yields Based on Machine Learning

Beena Pal[1], Prof. Priyanka Prajapati[2]
[1, 2] Department of Computer Science and Engineering
[1, 2] Alpine Institute of Technology, Ujjain, India

Abstract- *Crop yield prediction is a critically important forecasting problem trying to address food security in the world. This is a crucial AI based application where the previous crop yield is used to train a machine learning model and then the future crop yield is forecasted. The results can be used to decide, which crop to be sown during a particular season in a particular geographical location. Since the data to be analyzed is large, random and complex for analysis, hence conventional statistical techniques do not render high accuracy of prediction. This paper presents a review on the contemporary techniques for crop yield prediction.The performance evaluation parameters have also been discussed.. It is expected that the paper would present with a headway for further research in cloud workload prediction.*

Keywords- Cloud Crop Yield Predicition, Machine Learning, Artificial Neural Network (ANN),Mean Absolute Percentage error, Mean Square Error.

I. INTRODUCTION

Agriculture is arguable the most critical area of necessity as the entire population depends on agricultural produce. Currently, many countries are experiencing hunger because of the shortfall or absence of food with a growing population and critical pandemic situation. Expanding food production is a compelling process to annihilate famine. Developing food security and declining hunger by 2030 is a major objective for the United Nations. Hence crop protection and crop yield prediction are critically important to develop a sustainable agriculture environment throughout the world. It is an application of science for the benefit of the society.

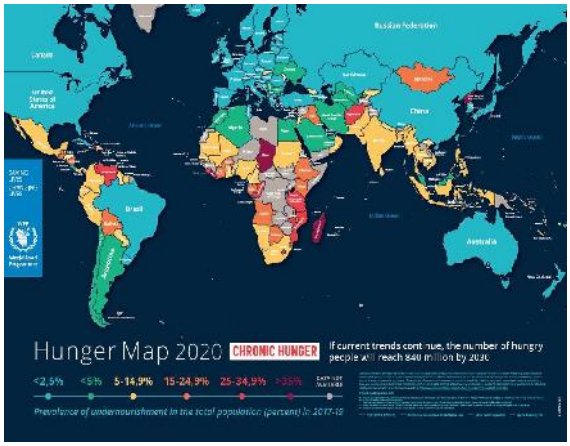

Fig.1 World Hunger Map

It is important for the accurate prediction of crop yields in agricultural countries, which can provide effective data support for national decision-makers, grain import, and export plans and solving food security issues. The vision of meeting world's food demands for the increasing population throughout the world is becoming more important in these recent years. Crop models and decision tools are increasingly used in agricultural field to improve production efficiency. The combination of advance technology and agriculture to improve the production of crop yield is becoming more interesting recently. Due to the rapid development of new higher technology, crop models and predictive tools might be expected to become a crucial element of precision agriculture. There are a lot of factors influencing crop productions, either directly or indirectly which affected the crop performance. Most of the factors that normally highlighted in researches are soil factors, such as pH, available nutrients, texture, organic matter content and soil-water relationships. There are also other factors highlighted in researched, e.g., weather and climatic factors including temperature, rainfall and light intensity; crop and cultivar; postharvest handling and storage; fertilizer applications and cultural practices [2]. These factors have formed a complex system for agriculture since it always deals with a large set of data for each problem it faced with. Agronomists and scientists still have doubt in finding a really suitable tool to review the available datasets based on current knowledge. Until today, a lot of approaches and modelling techniques have been in existence. Most of them represent the

data based on probabilities, which then are estimated by training and presented using algorithm by a human classifier [3]. Crop modelling was initially viewed as a tool to understand the performance of complex systems such as external factor like weather condition towards crop yield, at once helped us to understand the qualitative links between processed and crop performance [4]. For a number of years, the most crop models which were based on linear method were constructed from either linear or multiple linear regression or correlation analysis

II. ARTIFICIAL NEURAL NETWORKS

Artificial Intelligence and Machine Learning (AI &ML) are preferred techniques for analyzing large and complex data. Generally, artificial neural networks (ANN) are used for the implementation of artificial intelligence practically. The architecture of artificial intelligence can be practically implemented by designing artificial neural networks. The biological-mathematical counterpart of artificial neural networks has been shown below.

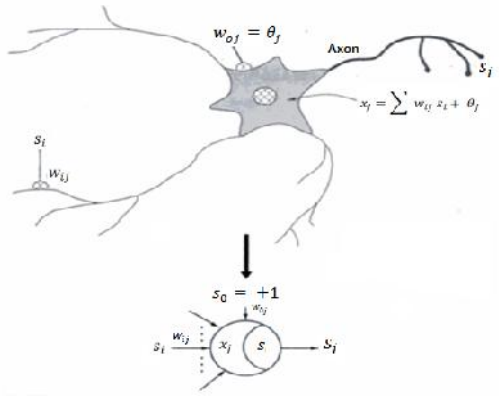

Fig.2 Biological-Mathematical Counterpart of ANN

The mathematical conversion of the ANN can be done by analyzing the biological structure of ANN. In the above example, the enunciated properties of the ANN that have been emphasized upon are:

1)Strength to process information in parallel way.
2) Learning and adapting weights
3)Searching for patterned sets in complex models of data.

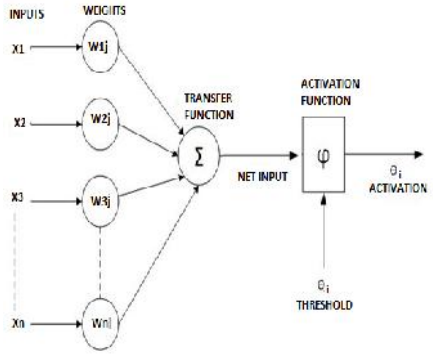

Fig.3 Mathematical Modeling of ANN

To see how the ANN really works, a mathematical model has been devised here, to indicate the functions mathematically.[7]. Here it is to be noted that the inputs of information parallel goes on into the input layer as specified whereas the end result analysis is marked from the output layer.

The feature of parallel acceptance and processing of data by the neural network serves a vital role. This ensures efficient and quicker mode of operation by the neural network. Also adding to it, the power to learn and adapt flexibly by the neural network aids in processing of data at a faster speed. [2]These great features and attributes make the ANN self dependent without requiring much intervention from humans. The output of the neural networks can be given by:

$$\boldsymbol{Y = \sum_{i=1}^{n} X_i . W_i + \theta_i} \quad \textbf{(1)}$$

Here,

Y represents output
X represents inputs
W represents weights
Θ represents Bias

Training of ANN is of major importance before it can be used to predict the outcome of the data inputs. Neural Networks can be used for a variety of different purposes such as pattern recognition in large and complex data pattern sets wherein the computation of parameters would be extremely daunting for conventional statistical techniques. The weights or the equivalents of experiences are evaluated and updated based on the data patterns which are fed to the neural networks for training. Thus using artificial neural networks for the prediction or forecasting of cloud workload is an efficient proposition.

III. PREVIOUS WORK

This section highlights the prominent work in the domain.

D. Elavarasan et al. in [1] proposed a a Deep Recurrent Q-

Network model which is a Recurrent Neural Network deep learning algorithm over the Q-Learning reinforcement learning algorithm to forecast the crop yield. The sequentially stacked layers of Recurrent Neural network is fed by the data parameters. The Q- learning network constructs a crop yield prediction environment based on the input parameters. A linear layer maps the Recurrent Neural Network output values to the Q-values. The reinforcement learning agent incorporates a combination of parametric features with the threshold that assist in predicting crop yield. Finally, the agent receives an aggregate score for the actions performed by minimizing the error and maximizing the forecast accuracy. The proposed model efficiently predicts the crop yield outperforming existing models by preserving the original data distribution with an accuracy of 93.7%.

C. Dang et al in [2] proposed a Redundancy Analysis (RDA) for feature selection. The simple effects of RDA were used to evaluate the interpretation rates of the explanatory factors. The conditional effects of RDA were adopted to select the features of the explanatory factors. Then, the autumn crop yield was divided into the training set and testing set with an 80/20 ratio, using Support Vector Regression (SVR), Random Forest Regression (RFR), and deep neural network (DNN) for the model, respectively. Finally, the coefficient of determination (R2), the root mean square error (RMSE), the mean absolute error (MAE), and the mean absolute percentage error (MAPE) were used to evaluate the performance of the model comprehensively. The results showed that the interpretation rates of the explanatory factors ranged from 54.3% to 85.0% ($p = 0.002$), which could reflect the autumn crop yields well. When a small number of sample training data (e.g., 80 samples) was used, the DNN model performed better than both SVR and RF models.\

Nigam et al. in [3] proposed a predicting the yield of the crop by applying various machine learning techniques. The outcome of these techniques is compared on the basis of mean absolute error. The prediction made by machine learning algorithms will help the farmers to decide which crop to grow to get the maximum yield by considering factors like temperature, rainfall, area, etc.

Y. Li et al. in [4] proposed a present our statistical modeling practices for predicting rainfed corn yield in the Midwest U.S. and address the aforementioned issues through comprehensive diagnostic analysis. Our results show that vapor pressure deficit and precipitation at a monthly scale, in spline form with customized knots, define the “Best Climate-only” model among alternative climate variables (e.g., air temperature) and fitting functions (e.g., linear or polynomial), with an out-of-sample (leave-one-year-out) median R2 of 0.79 and RMSE of 1.04 t/ha (16.6 bu/acre) from 2003 to 2016.

Bhosale et al. in [5] proposed a data mining approach on agricultural data for yield prediction. It would help in getting results using technologies like data analytics and machine learning and this result will be given to farmers for better crop yield in terms of efficiency and productivity. Authors studied K-means clustering which was used to create clusters. Data stored in clusters facilitated fast search in less time based on cluster hypothesis. The work is expected to help farmers to increase the yield of their crops. Storage of big data in clusters by using K-means clustering algorithm, reduce it to appropriate/valid content using the algorithm. Apriori algorithm helped to count frequently occurring features which helped to predict crop yield for specific location.

T. Islam et al. in [6] proposed an Artificial Neural Network based approach for modeling and prediction of crop yield. This algorithm aims to get better output and prediction, as well as, support vector machine, Logistic Regression, and random forest algorithm is also considered in this study for comparing the accuracy and error rate. Moreover, all of these algorithms used here are just to see how well they performed for a dataset which is over 0.3 million. We have collected 46 parameters such as - maximum and minimum temperature, average rainfall, humidity, climate, weather, and types of land, types of chemical fertilizer, types of soil, soil structure, soil composition, soil moisture, soil consistency, soil reaction and soil texture for applying into this prediction process.

Fernandez et al. in [7] proposed a techniques for the he estimation of yield and total volume of maize production using Spot-5 satellite images and empirical models. These models expressed a) yield (Y) as a function of LAI, and b) yield as a function of NDVI. To determine the efficiency degree of the calculated predictions at the flowering stage of the crop, yield sampling was done at the physiological maturity stage in pilot plots. Regarding yield prediction in the flowering stage, the models Y = f (LAI) reported a value of 5.96 ton.ha-1 and the model Y = f (NDVI) a value of 5.04 ton.ha-1 was obtained. These data represent 114% and 97% respectively of the true yield recorded on the field. The models are specific to the maize crop and the cultivated plots location, and that the forecasts can be acceptably accurate provided the sown areas are precisely determined.

Huang et al. in [8] present a Bayesian model averaging (BMA) method for a multiple crop-growth model ensemble to provide more reliable predictions of maize yields in Liaoning Province, China. The integrated prediction is achieved using a linear combination of the three ensemble members using BMA weights. This integrated approach results in more

accurate and precise predictions than any individual model over the entire province. This is because the BMA framework effectively compensates for the uncertainty of individual model simulation and takes advantage of each competing model for reliable prediction.

N. Gandhi et al. in [9] proposed a prediction model based on machine learning. The parameters considered for the present study were precipitation, minimum temperature, average temperature, maximum temperature and reference crop evapotranspiration, area, production and yield for the Kharif season (June to November) for the years 1998 to 2002. The dataset was processed using WEKA tool. A Multilayer Perceptron Neural Network was developed. Cross validation method was used to validate the data. The results showed the accuracy of 97.5% with a sensitivity of 96.3 and specificity of 98.1. Further, mean absolute error, root mean squared error, relative absolute error and root relative squared error were calculated for the present study. The study dataset was also executed using Knowledge Flow of the WEKA tool. The performance of the classifier is visually summarized using ROC curve.

P. Bose et al. in [10] proposed a spiking neural networks (SNNs) for remote sensing spatiotemporal analysis of image time series, which make use of the highly parallel and low-power-consuming neuromorphic hardware platforms possible. This paper illustrates this concept with the introduction of the first SNN computational model for crop yield estimation from normalized difference vegetation index image time series. It presents the development and testing of a methodological framework which utilizes the spatial accumulation of time series of Moderate Resolution Imaging Spectroradiometer 250-m resolution data and historical crop yield data to train an SNN to make timely prediction of crop yield. The research work also includes an analysis on the optimum number of features needed to optimize the results from our experimental data set. The proposed approach was applied to estimate the winter wheat (Triticum aestivum L.) yield in Shandong province, one of the main winter-wheat-growing regions of China.

IV RESEARCH GAP IDENTIFIED AND FUTURE SCOPE

1) There may be local fluctuations in the data which need to be filtered out to train a system. The selection of an appropriate mathematical tool in this regard is important.
2) Limited work has been done on use of advanced data pre-processing techniques which can lead to higher predictive performances of classifier with lower computational effort.

The future scope for research should be the use of data optimization techniques along with feedback mechanism for the errors to propagate in each iteration so as to attain the following objectives:

1. Faster convergence
2. Lesser errors in each iteration successively.

V. PERFORMANCE METRICS

Since the purpose of the proposed work is time series prediction, hence it is necessary to compute the required performance metrics. Since there is a chance of positive and negative errors to cancel out, hence it is necessary to compute the Mean Absolute Percentage Error (MAPE) given by:

$$MAPE = \frac{100}{M}\sum_{t=1}^{N}\frac{E-E_t}{E_t} \quad (2)$$

Here,
N is the total number of samples
E is the actual value
$\mathbf{E_t}$ is the predicated value

The mean square error is also evaluated often to stop training, which is given mathematically by:

$$MSE = \frac{1}{N}e_i^2 \quad (3)$$

Here,
E is the error
N is the number of samples

It is always envisaged to attain low error values and high values of accuracy for cloud workload prediction.

VI. CONCLUSION

The present work renders insight into the basic methodologies working as empirical models for crop yield prediction, as a time series prediction.It has been shown that crop yield prediction is a critically important forecasting problem trying to address food security in the world. This is a crucial AI based application where the previous crop yield is used to train a machine learning model and then the future crop yield is forecasted. The results can be used to decide, which crop to be sown during a particular season in a particular geographical location, as well as bolster government agricultural policies. The paper summarizes the prominent work done in the domain and its salient points.

REFERENCES

[1] D. Elavarasan and P. M. D. Vincent, "Crop Yield Prediction Using Deep Reinforcement Learning Model for Sustainable Agrarian Applications," in IEEE Access, vol. 8, pp. 86886-86901, 2020, doi: 10.1109/ACCESS.2020.2992480.
[2] C Dang, Y Liu, H Yue, JX Qian, “Autumn Crop Yield Prediction using Data-Driven Approaches:-Support

Vector Machines, Random Forest, and Deep Neural Network Methods", Canadian Journal of Remote Sensing, Taylor and Franscis, 2020.

[3] A. Nigam, S. Garg, A. Agrawal and P. Agrawal, "Crop Yield Prediction Using Machine Learning Algorithms," 2019 Fifth International Conference on Image Information Processing (ICIIP), 2019, pp. 125-130, doi: 10.1109/ICIIP47207.2019.8985951.

[4] Y Li, K Guan, A Yu, B Peng, L Zhao, B Li, J Peng, "Toward building a transparent statistical model for improving crop yield prediction: Modeling rainfed corn in the US", Field Crops Research Journal, Elsevier 2019, vol. 234, Issue-15 pp. 55-65

[5] S. V. Bhosale, R. A. Thombare, P. G. Dhemey and A. N. Chaudhari, "Crop Yield Prediction Using Data Analytics and Hybrid Approach," 2018 Fourth International Conference on Computing Communication Control and Automation (ICCUBEA), 2018, pp. 1-5, doi: 10.1109/ICCUBEA.2018.8697806.

[6] T. Islam, T. A. Chisty and A. Chakrabarty, "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh," 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), 2018, pp. 1-6, doi: 10.1109/R10-HTC.2018.8629828.

[7] Y. M. Fernandez-Ordoñez and J. Soria-Ruiz, "Maize crop yield estimation with remote sensing and empirical models," 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), 2017, pp. 3035-3038, doi: 10.1109/IGARSS.2017.8127638.

[8] X Huang, G Huang, C Yu, S Ni, L Yu, "A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging", Journal of Field Crops Research, Elsevier 2017. Vol. 211, pp. 114-124

[9] N. Gandhi, O. Petkar and L. J. Armstrong, "Rice crop yield prediction using artificial neural networks," 2016 IEEE Technological Innovations in ICT for Agriculture and Rural Development (TIAR), 2016, pp. 105-110, doi: 10.1109/TIAR.2016.7801222.

[10] P. Bose, N. K. Kasabov, L. Bruzzone and R. N. Hartono, "Spiking Neural Networks for Crop Yield Estimation Based on Spatiotemporal Analysis of Image Time Series," in IEEE Transactions on Geoscience and Remote Sensing, vol. 54, no. 11, pp. 6563-6573, Nov. 2016, doi: 10.1109/TGRS.2016.2586602.

[11] T. U. Rehman, S. Mahmud, Y. K. Chang, J. Jin, and J. Shin, ''Current and future applications of statistical machine learning algorithms for agricultural machine vision systems,'' Comput. Electron. Agricult., vol. 156, pp. 585–605, Jan. 2019.

[12] D.Elavarasan,D.R.Vincent,V.Sharma,A.Y.Zomaya,andK. Srinivasan, ''Forecasting yield by integrating agrarian factors and machine learning models: A survey,'' Comput. Electron. Agricult., vol. 155, pp. 257–282, Dec. 2018.

[13] M. D. Johnson, W. W. Hsieh, A. J. Cannon, A. Davidson, and F. Bédard, ''Cropyield forecasting on the Canadian Prairies by remotely sensedvegetation indices and machine learning methods,'' Agricult. Forest Meteorol., vols. 218–219, pp. 74–84, Mar. 2016.

[14] [14] A. Kaya, A. S. Keceli, C. Catal, H. Y. Yalic, H. Temucin, and B. Tekinerdogan, ''Analysis of transfer learning for deep neural network based plant classification models,'' Comput. Electron. Agricult., vol. 158, pp. 20–29, Mar. 2019.

[15] A. Kamilaris and F. X. Prenafeta-Boldú, ''Deep learning in agriculture: A survey,'' Comput. Electron. Agricult., vol. 147, pp. 70–90, Apr. 2018.

[16] I.M.Evans,''Reinforcement,principle,''inInternationalEncyclopediaof the Social & Behavioral Sciences, J. D. Wright, 2nd ed. Amsterdam, The Netherlands: Elsevier, 2015, pp. 207–210.

[17] D.VogiatzisandA.Stafylopatis,''Reinforcementlearningfor ruleextraction from a labeled dataset,'' Cognit. Syst. Res., vol. 3, no. 2, pp. 237–253, Jun. 2002.

[18] S. Wan, Z. Gu, and Q. Ni, ''Cognitive computing and wireless communications on the edge for healthcare service robots,'' Comput. Commun., vol. 149, pp. 99–106, Jan. 2020.

[19] A.Tolba,O.Said,andZ.Al,Makhadmeh,''MDS:Multi-leveldecisionsystem for patient behavior analysis based on wearable device information,'' Comput. Commun., vol. 147, pp. 180–187, Nov. 2019.