CROP YIELD PREDICTION USING DEEP LEARNING

Mrs. Mamatha K, Shantideepa Samanta, Kundan Kumar Prasad

Don Bosco Institute of Technology, Bangalore, India

mamathaise@dbit.co.in, samanta.shantideepa@gmail.com, prasadkundan2019@gmail.com

Abstract- Agriculture provides a living for around 58 percent of India's population. Agriculture, forestry, and fisheries were expected to generate ₹19.48 lakh crore in FY20. Given the significance of agriculture in India, farmers might benefit from early forecasting of agricultural yields. The study focuses on predicting agricultural yield,, for Karnataka state using the regression with neural network model. The final constructed dataset takes parameters like agricultural area, crop, taluka, year, season, district wise annual rainfall (mm), district wise maximum and minimum temperature (˚C) and harvest or yield for the time period of 1997 to 2017. The underlying model is built utilizing a Multilayer Perceptron Neural Network, a ReLu Activation function, an Adam Optimizer, and 50 epochs with a batch size of 200. The end of the training gained 96.43% accuracy on test data. Several additional well-known regression algorithms such as Multinomial Linear Regression, Random Forest Regression and Support Vector Machine are also constructed and trained using the same dataset so as to compare their performance to the base model. From the final comparison results it was found that neural network model has outperformed classic machine models for crop yield prediction in terms of both mean absolute error and accuracy.

Keywords: Neural Network, Support Vector Regression, Random Forest Regression, Linear Regression, ReLu, Adam Optimizer

INTRODUCTION

Agriculture in India stretches back to the Indus Valley Civilization Era, and maybe much earlier in some regions of Southern India. In terms of agriculture output, India stands second in the world. While agriculture's proportion of the Indian economy has gradually decreased according to the fast development of the industrial and service sectors, to less than 15%, the sector's relevance in India's social and economic fabric stretches much beyond this statistic. The reason for this deterioration in the agriculture industry is because farmers are not empowered, and there is a lack of application of information technology in the farming sector. Farmers are less knowledgeable about the crops they cultivate. We typically overcome this challenge by utilizing appropriate deep learning algorithms to forecast crop output and name based on a variety of parameters like as temperature, rainfall, season, and location. Based on the data source supplied by the Indian government, this study presents a Neural Network system to forecast agricultural production and crop success rate. The main challenge encountered when assembling the work was the lack of a single source dataset to train the suggested model on. To address these issues, all dispersed data is collected and relevant feature engineering and data pre-processing steps are employed. The data source is massive, comprising of records for all the areas of India that were sieved and processed to acquire records for Karnataka state, resulting in 12961 entries. The cycle of the crop data for summer, Kharif, Rabi, fall, and the entire year is used. To obtain records for the state of Karnataka, the dataset is pre-processed using Pandas and Profiling tools of pandas in Python. The crop yield forecast model employs an artificial neural network's back propagation technique. The technology of multilayer perceptrons is employed. The proposed work has a wide range of applications in improving real-world farming conditions. Every year, a large amount of crop is damaged owing to a lack of

understanding of weather patterns such as temperature, rainfall, and so on, which have a significant impact on crop output. This initiative not only aids in forecasting these characteristics throughout the year, but it also aids in projecting agricultural yields in various seasons based on historical trends. As a result, it enables farmers to select the best crop to plant in order to incur the fewest losses. Different regression models are also constructed using machine learning, and their efficiency and accuracy are compared to the Neural Network model in order to provide some tangible results.

RELATED WORK

This study examines the rice productivity and sustainability producing regions that are reliant on acceptable climatic conditions by deducing experimental findings acquired by using an SMO classifier employing the WEKA tool to a dataset of 27 districts in the state of Maharashtra.[1]. The suggested work uses soil and PH samples as input and predicts crops that are suited for the soil and fertilizer that can be utilized as a solution in the form of a webpage. So, soil information is gathered using sensors, and the data is transferred from the Arduino through Zigbee and WSN (Wireless Sensor Network) to MATLAB, where it is analyzed and processed using ANN (Artificial Neural Network), and crop suggestions are made using SVM.[2]. The major aim for this research is to look at the linkages between the climate and the yield of a crop and the historical trends and changes that has an impact in climate-related factors on finger millets, rice, wheat and maize production in Nepal's Morang region. [3]. This study intended to evaluate these novel data mining tools and finger millet finger millet finger millet tools and mechanisms and enroll them to the database's numerous characteristics to see whether relevant correlations could be discovered. [4]. The major goal of this research was to test a strategy for incorporating satellite imaging characteristics into A crop development methodology was built to approximate the wheat yields during spring season at the macro level and computed separately. [5]. This study forecasts the suitable crop using crop yield prediction algorithms that detect distinct soil characteristics and meteorological condition factors. It illustrates the artificial neural network algorithm's capacity to monitor and predict agricultural production in remote and rural locations.[6]. The goal of this study is to conduct a comprehensive assessment of the current literature on DL methods in BDA. The findings might aid researchers in developing double-armed prognostic studies that properly evaluate the potential of the major elements of Deep Learning techniques, study the effectiveness of Deep Learning approaches in increasing Big Data Analysis, and investigate the benefits and drawbacks of Deep Learning approaches..[7]. In this study, researchers used a Multi-Layer Perceptron, back propagation based on feed forward deep neural network to forecast wheat production (ANN). [8]. The suggested model is an example of a more sophisticated model that uses a highly complex set of variables to forecast wheat production. ECOSYS and SIRIUS, like CERES, are sophisticated models that integrate a multitude of factors and heavily rely on computer modeling to anticipate wheat growth. [9]

PROPOSED IDEA

Figure 1 depicts the modules constructed in the proposed work. It is made up of three major components. One module is dedicated to early yield forecast, utilizing characteristics such as crop area, yearly rainfall, temperature data, and Karnataka state output history from 1998 to 2014. The second module is the fertilizer recommendation system, which takes into account the quantity of three major nutrients in the soil, namely nitrogen, phosphorus, and potassium, as well as the crop to be sown, and suggests the fertilizer that may be utilized to increase crop output. The third module is a crop-specific WPI trends indicator that depicts the whole index price over the following 12 months graphically.

DATASET

The data for this study was obtained from the Indian government's website. The datasets are freely accessible for research and scholarly purposes. The collection contains information spanning the years 1997 to 2017. All the required datasets are collected and pre-processed to obtain a final dataset which will be used to train the model. For the experiment in this investigation, the following parameters are used.

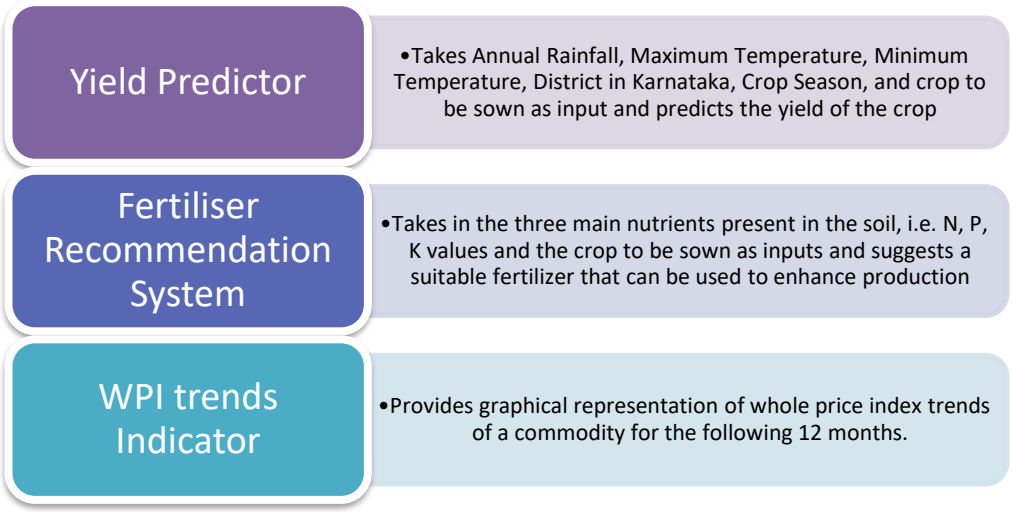

Figure 1. System Modules

- **Crop** – Rice, Wheat, Castor seeds, Bajra, Arhar, Season, groundnut, cottonseed, tur, and other crops are included in the dataset.
- **State**-Karnataka
- **District** – 'BAGALKOT', 'BENGALURU RURAL', 'BELLARY', 'BELGAUM', 'CHIKMAGALUR', 'CHITHADURGA', 'DHARWAD', etc
- **Season**- Autumn, Summer(Kharif), Winter(Rabi), and Yearly
- **Year-** 1997 to 2017
- **Rainfall** - Monthly rainfall data (mm) for each district of Karnataka State, whose sum is taken to evaluate annual rainfall and concatenated to the final dataset.
- **Temperature** - District wise maximum and minimum temperature (˚C), who's mean is calculated and appended to the final dataset
- **Production** - It is expressed as tonnes per hector in million.
- **Fertilizers** - Describes the amount of N, P, K required in the soil in order to grow a specific crop in a region.

The data in the government dataset has been checked for outliers and anomalies. The parameters were also translated to numerical and category formats to meet the model's requirements. Figure 2 displays a bar graph from 1998 to 2014 that correlates the season and produce of all Karnataka districts. According to the graph below, the bulk of the crops cultivated in Karnataka are year-round crops.

METHODOLOGY

To integrate all of the data sets obtained for this investigation, Microsoft Office Excel was employed.

Step 1: Acquiring monthly average records from Indian Government databases for each parameter (rainfall, minimum, median, maximum temperature, and reference crop transpiration) from 1997 to 2017.

Step 2: Calculate the total precipitation for each region during the summer (Kharif) season each year (July to November).

Step 3: Computing the average temperature for each region all throughout summer (Kharif) season for the minimum, average, and maximum temperatures annually.

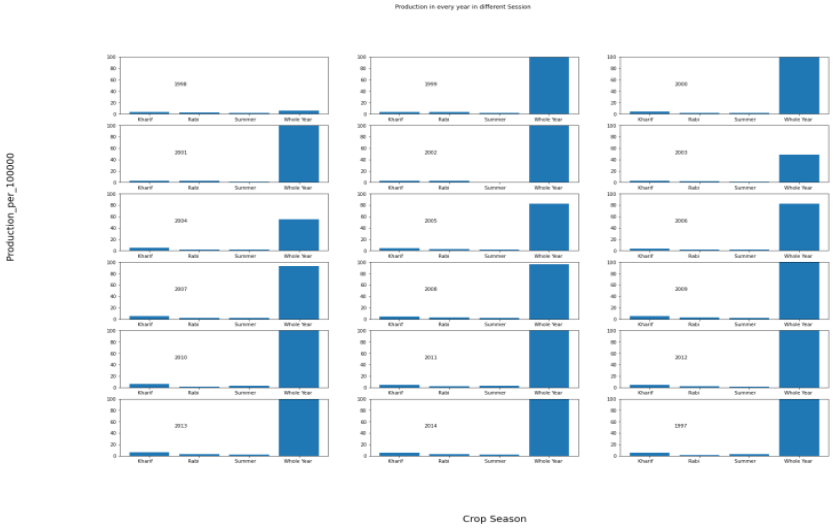

Figure 2. Year-specific Bar graph to depict the correlation of season vs. Production

Step 4: During the Kharif season, calculate the average reference crop evapo-transpiration for each year for each district in Karnataka state.

Step 5: Obtaining information on each district's acreage, output, and rice crop yield from 1997 to 2017, based on publicly accessible Indian Government sources and databases.

Step 6: The raw collected data was then compiled in Excel Spreadsheet into a single document with the following columns: name of the region, year, rainfall, minimal temperature, average temperature, maximum temperature, area, production, and yield.

Step 7: Because specific climatic characteristics of a year or harvest quantity statistics were not accessible for some of the districts, those records were removed. The data from that specific year was not included in the current study. Each record now has a record number.

Step 8: Unrequired columns were deleted from the data set in order to prepare it for use with the multilayer perception method. They were name of the district, and year.

Step 9: All the records was then arranged according to area. The current study did not take into account areas less than 100 hectares. As a result, those records were eliminated

Step 10: Using sklearn, the attribute values in the label is transformed to encoding.

Step 11: The entire data source was then sorted based on the date of production.

Step 12: The study then evaluates harvest yield as an output parameter, as well as characteristics such as type of crop, cultivation area, taluk, and season.

Step 13: The final dataset was then sorted and compiled into a.csv file for iterative use in Python Tensor Flow with the multilayer perceptron method.

Step 14: Model is trained in this step. Using linear regression with a neural network with three layers and the Adam optimizer.

Step 15: Using an 80:20 split, the data source was segregated into test and train set.

DATA FLOW DIAGRAM

The diagram below depicts the flow of data through the system. The flow of all modules stays constant, with the only variation being the final result. Inputs for the relevant modules, such as yearly rainfall, temperature, district, crop name, season, and fertiliser data, are obtained via a web-based application by the user. A JSON data object is returned, which has been scaled with the sklearn package. The categorical data, such as district, season, and crop name, is again one hot encoded, and the data object is ultimately transformed to a numpy array. This information is subsequently put into the Neural Network model.

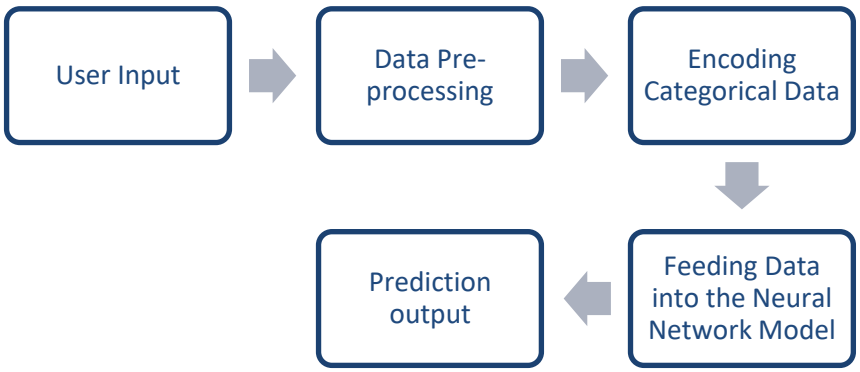

Figure 3. Dataflow Diagram of the proposed System

ARTIFICIAL NEURAL NETWORK- AN OVERVIEW

An Artificial Neuron is essentially a biological neuron engineering technique. It has a gadget with several inputs and just one output. ANN is made up of a huge number of basic processing units that are linked and stacked. [10]. An ANN starts with a phase of training in which It begins to identify patterns in data, whether they are visual, audio, or linguistic. During this supervised phase, the network compares its expected results to what it was supposed to produce—the projected output. Back propagation is employed to reconcile the disparity between the two findings. This implies that the system operates backward, from the output unit to the input units, adjusting the weight of its links between the modules until the gap between the estimated and planned results produces the smallest feasible error.

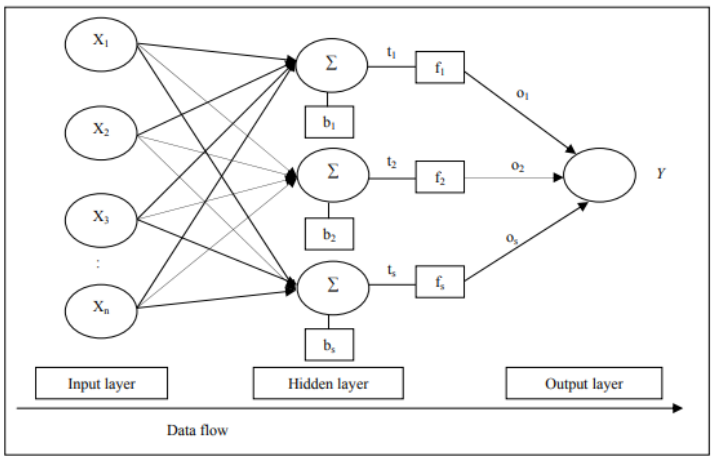

Figure 4.Layers and connections of ANN model

TRAINING SPECIFICATIONS

The following are the important parameters that were evaluated for testing

- Layer : 3
- Neuron at each layer : Layer 1, Layer 2 = 20
- Layer 3 = 1
- batch_size=100,
- Activation : ReLu
- Oprimizer = Adam
- epochs=50
- kernel_initializer='uniform
- lr rate : 0.01

PERFORMANCE EVALUATION

We utilize scatter plots to compare the actual test data output to the predictions generated by the model on the test data. The graph below illustrates a linear connection between the actual and predicted results. A positive slope with a strong correlation between the actual and anticipated results indicates a greater success rate. It can also be stated that for the majority of the test inputs, the model was able to forecast a yield with extremely low error to the actual yield output.

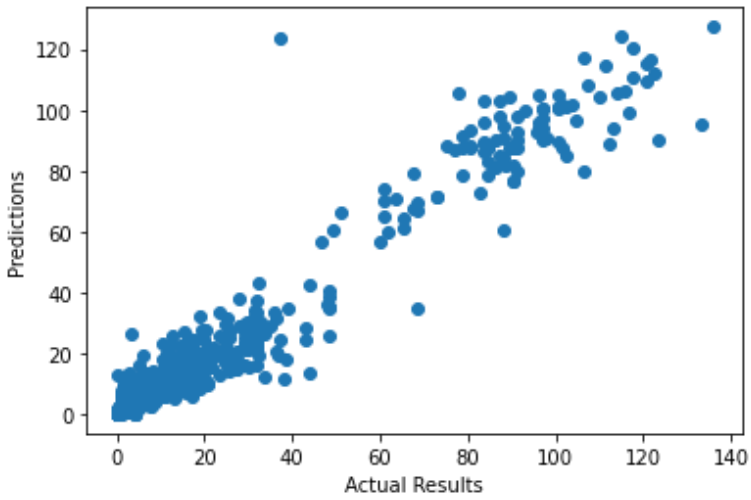

Figure 4.Linear Correlation between actual and predicted results

The algorithm's performance is measured using the two metrics listed below.

MEAN ABSOLUTE ERROR:

MAE is a weighted average magnitude of absolute differences between N estimated vectors S = x1, x2,..., xN and S_ = y1, y2,..., yN, with the associated loss function calculated as:

$$L_{MAE}(S, S^*) = \frac{1}{N} \sum_{i=1}^{N} || \ x_i - y_i \ ||$$

where || · || denotes L1 norm. [11].

R-SQUARED:

R-squared (R2) is a quantitative measure that reflects the percentage of a dependent variable's variance explained by an independent variable or variables in a regression model. Whereas correlation shows the significance of the association between an independent and dependent variable, R-squared represents how well the fluctuation of one variable explains the variation of the other. The fraction of a fund's or security's movements that can be accounted by changes in a standard deviation is typically described as R-squared. The R-Squared score for the proposed work was able to reach approximately 0.9645 within 50 epochs. The score can be calculated using the formula below:

$$R^2 = 1 - \frac{RSS}{TSS}$$

Where,

R^2 is coefficient of determination

RSS is Sum of Squares of residual

TSS is total sum of squares

The graph below depicts the metric scores discussed above for our neural network model.

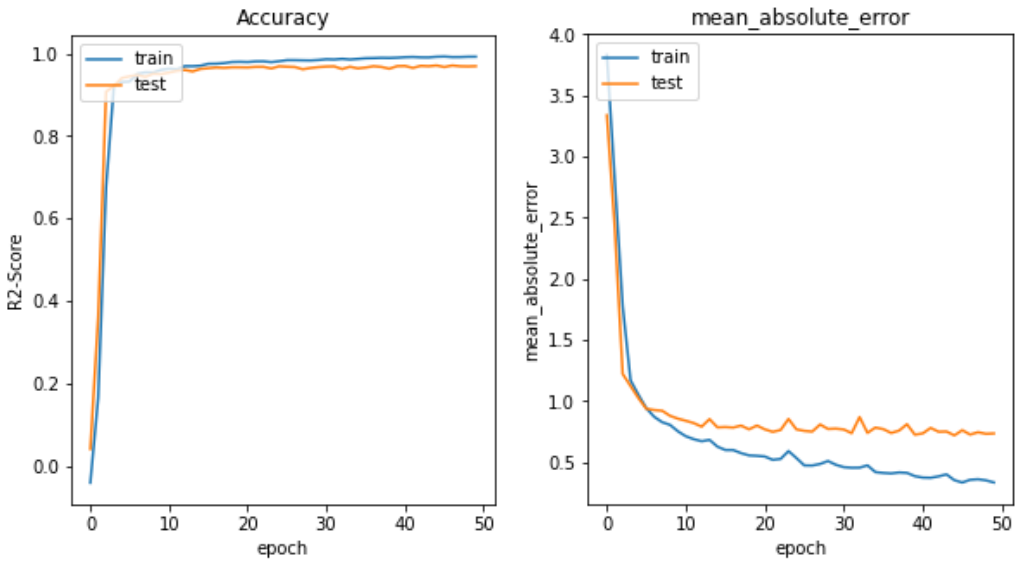

Figure 5. Performance graphs of the model

EXPERIMENTAL RESULTS

In the current study, we collected multiple datasets and performed appropriate feature engineering to build a single source of data that accounts for all of the essential features to help model correctness.

To provide a comparison study of our neural network model's performance, we utilize the same dataset to train three additional regression models, namely, the Multinomial Regression model, the Random Forest regression model, and the support vector regression model. All three models' performance was evaluated using the same two measures described above: mean absolute error and R-Squared score.

The bar graph below compares the performance of all three models, as well as our targeted model, the neural network model. It is clearly shown that the Artificial Neural Network outperforms the other regression methods in terms of accuracy and error rate.

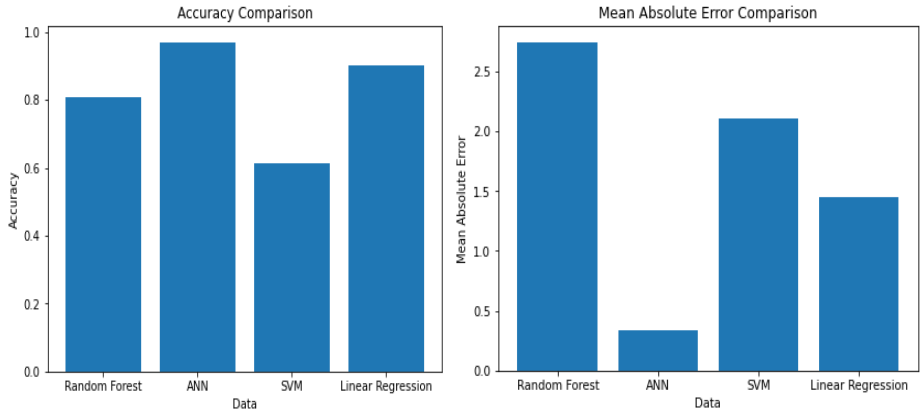

Figure 6. Comparative Analysis of the performance of all the regression models

DISCUSSIONS AND CONCLUSIONS

A non-linear way of interpreting the connection is necessary to demonstrate the interactions between the factors impacting crop production. Because of the complexities of the elements influencing crop output, a linear approach such as linear regression was judged insufficient to depict the interconnections between the factors and crop yield. For forecasting agricultural yield, ANN was thought to be a viable alternative to standard regression methods. A neural network not only predicts non-linear correlation successfully, but it can also recognize complicated patterns in data and train appropriately, something most traditional approaches fail to do.

REFERENCES

[1]. Niketa Gandhi, Leisa J. Armstrong, Owaiz Petkar, Amiya Tripathi, "Rice Crop Yield Prediction in India using Support Vector Machines", 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)

[2]. Preethi G, Rathi Priya V, Sanjula S M, Lalitha S D, Vijaya Bindhu B, "Agro based crop and fertilizer recommendation system using machine learning", European Journal of Molecular & Clinical Medicine, Volume 7, Issue 4, 2020, ISSN: 2515-8260

[3]. Badri Khanal, "CORRELATION OF CLIMATIC FACTORS WITH CEREAL CROPS YIELD: A STUDY FROM HISTORICAL DATA OF MORANG DISTRICT, NEPAL", The Journal of Agriculture and Environment Vol: 16, June 2015

[4]. Raorane A.A., Kulkarni R.V., "Review- Role of Data Mining in Agriculture", (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 270 - 272, ISSN: 0975-964

[5]. Paul C. Doraiswamy, Sophie Moulin, Paul W. Cook, and Alan Stern, "Crop Yield Assessment from Remote Sensing", Photogrammetric Engineering & Remote Sensing Vol. 69, No. 6, June 2003, pp. 665–674.

[6]. Teresa Priyanka, Pratishtha Soni, C. Malathy , "Agricultural Crop Yield Prediction Using Artificial Intelligence and Satellite Imagery", Eurasian Journal of Analytical Chemistry, 2018 13 (SP): 6-12, ISSN: 1306-3057

[7]. Hordri N. F., Samar, A., Yuhaniz S. S., Shamsuddin S. M., "A Systematic Literature Review on Features of Deep Learning in Big Data Analytics", Int. J. Advance Soft Compu. Appl, Vol. 9, No. 1, March 2017, ISSN: 2074-8523

[8]. Muhd Khairulzaman Abdul Kadir, Mohd Zaki Ayob, Nadaraj Miniappan, "Wheat Yield Prediction: Artificial Neural Network based Approach", 2014 4th International Conference on Engineering Technology and Technopreneuship (ICE2T).

[9]. R. J. Brooks, et al., "Simplifying Sirius: sensitivity analysis and development of a meta-model for wheat yield prediction," European Journal of Agronomy, vol. 14, pp. 43-60, 2001.

[10]. Manish Mishra, Monika Srivastava, "A View of Artificial Neural Network", IEEE International Conference on Advances in Engineering & Technology Research (ICAETR - 2014), August 01-02, 2014

[11]. Jun Qi, Jun Du, Sabato Marco Siniscalchi, Xiaoli Ma, Chin-Hui Lee, "On Mean Absolute Error for Deep Neural Network Based Vector-to-Vector Regression", IEEE SIGNAL PROCESSING LETTERS, VOL. 27, 2020