Deep Neural Network Model for Proficient Crop Yield Prediction

K. Pravallika¹ G. Karuna¹*, K. Anuradha¹, V. Srilakshmi¹

¹ Computer Science and Engineering, GRIET, Hyderabad, Telangana, India.

Abstract. Crop yield forecasting mainly focus on the domain of agriculture research which has a great impact on making decisions like import-export, pricing and distribution of respective crops. Accurate predictions with well-timed forecasts is very important and is a tremendously challenging task due to numerous complex factors. Mainly crops like wheat, rice, peas, pulses, sugarcane, tea, cotton, green houses etc. can be used for crop yield prediction. Climatic changes and unpredictability influence mainly on crop production and maintenance. Forecasting crop yield well before harvest time can help farmers for selling and storage. Agriculture deals with large datasets and knowledge process. Many techniques are here to predict the crop yield. Farmers are benefited commercially by these predictions. Factors such as Genotype, Environment, Climatic conditions and Soil types used in predicting the Yield. For predicting accurately we need to know the fundamental understanding and relationship between the interactive factors and the yield to reveal the relationships between the datasets which are comprehensive and powerful algorithms. Based on the study of various survey papers it has been found that in all the crop predictions, various deep learning, machine learning and ANN algorithms implemented to predict yield forecast and the results are analyzed.

1. Introduction

Yield forecasting of Crop is very essential for the production of food globally. Accurate forecasting can be done by policy makers to take timely decisions to import and export in terms of improving the national food security. Seed companies must forecast new hybrids in order to breed better varieties for various types of environments. Financial decisions can be easily taken and get benefitted by farmers and Growers. Genotype and Environmental interactions are highly complex characteristics. Genotype means the genetic characteristic of individual for particular trait and phenotype of individual deals with observable characteristics like physical, physiological, biochemical and behavioral. For example, length of the plant is one type of phenotype trait.

High dimensional marker data typically contains millions of markers for each plant individual and is represented by genotype.

*Corresponding author: karunavenkatg@gmail.com

The impact of genetic markers must be estimated, which results in interactions with various environmental factors and field management activities that must be calculated.

Fig.1. Genotype, phenotype and environmental interactions

Alleles is the alternative word for genes and each individual inherits two alleles for a character one from male and one from female. Two alleles can be same (TT or tt) or different (Tt).
Fig. 2. Alleles, genes and traits

Dominant allele that expresses itself and recessive allele unable to express itself. For a particular individual if the dominant allele and recessive allele is known then prediction of phenotype will become simple.

Environmental factors will have great effect on genotype. One of the explicit functions is phenotype, which consists of genotype is treated as G, E as environment, and their interactions are considered as G*E and is treated as noise.

Fig. 3. G*E interactions

Group of environments will share same type of varieties can be considered as mega environments. Kharif (july-oct) and Rabi(mar-june) are the two main seasons for growing different crops. Rice, maize, sorghum, bajra, ragi, pulses, soyabean, groundnut and cotton are the crops grown in kharif season.

Wheat, barley, oats, chick pea, mustard seeds are grown in rabi season. Syngenta is one of the company which provides the data for crop predictions. Climate data, which encompasses length of day, rainfall patterns, radiation from the sun, air humidity, min and max temperatures, is indeed one of the factors.

2. Related Work

In the Region of maharashtra, algorithms are used to predict paddy crop yields using anns[2]. Maharashtra's 27 districts have datasets inferred from available public Indian government records. When predicting, snowfall, lower limit temperature, mean temperature, highest temperature, region, output, and harvest are all factored into the equation. Weka tool is used for processing the dataset. In this rice crop prediction a multilayer perceptron neural network and is better than the regression models. Multilayer perceptron, radial basis function networks, and kohonen self organising feature maps are used in neural network[17] models.

Machine learning approaches are used to predict crop yields, and they are implemented in the PHP platform. Temperature, area, humidity, and production are used to predict tomato crop parameters. Ch.Vishnuvardhan Chowdary, Dr.K.Venkataramana[3] developed id3 algorithm for quality crop yield prediction. Different types of fertilizers are used to increase fertility of soil by adding nitrogen nutrients. Pesticides are used to remove the pest from the crop. Gradient boosting regressor uses cross validation to get more accuracy i.e 87.9%. Random forest regressor gets the accuracy of 98.9%.

Crop yield prediction develops a mobile application design and implementation using machine learning[4], it mainly helps the farmers to predict production of specific crop in particular regions. Rainfall and temperature are the physical parameters used in predicting the yield of crop. The app is intended to help farmers by requiring them to enter information such as location, farm size, size, temperature, rainfall, and crop dataset. To train the dataset ARIMA model is used. Accurate predicton can be done based on the availability of quality data. Python, data processing, and Android Studio are used to design the application. Intelligrow is the smart phone application in which the results are portrayed in systematic manner.

The learning algorithm to forecasting banana cultivation yields[5] utilises long - short - term memory strands with various perceptrons. Information is derived from beneficiaries (ARB) of Dapco in Davao del Norte, Philippines. Epoch, batch size, neurons are used as model parameters to identify the optimal values. Floods and droughts are the extreme events. Banana is the tropical fruit crop in philippines.
In order to derive and refine yield prediction data, a review of research for estimation using machine learning[6] was initiated. Six databases were searched to find 567 related research, of which 50 have been chosen for some further exploration based on selected studies. Air temp, soil moisture, and soil composition are all used by artificial intelligence. Popular deep neural networks are CNN, LSTM, and DNN. Machine learning historical data is used for training phase and testing phase for performance evaluation. The literature review and study objectives serve as basis for descriptive and inferential machine learning algorithms. Models of description allow for the acquisition of knowledge from collected data. Predicting future result can be done for predictive models. Systematic literature review(SLR) gives the overview of the crop yield[13] prediction and its result is different from other results. Objectivity and transparency are the SLR factors.

An ANN-based weather prediction model [7] asserts a linear correlation between both the weather data entry and the known target data. Atmosphere and state of given location is predicted in weather forecasting which collects the quantitative data. To protect life and property weather warnings are the important forecasts. ANN mainly minimizes the error. Temperature forecasts have special interest on minimum and maximum temperature of day. For winter, summer and fall training and testing is done separately. RBFN and ensembles outperformed all single networks.

Recursive neural network (BPN) [8] has the key advantage of being comparatively imprecise for a multitude of tasks. A multilayer perceptron has 3 levels: the input nodes, the intermediate hidden nodes, and the output units. Factor analytic models are used to join environments and genotypes that do not have cross-over genotype * environment interactions[9]. Prediction assessment of environmental trains[11] using linear mixed models[18]. Syngenta datasets are used for predicting the yield of the crop across different countries. Crop yield prediction using deep neural networks[1] uses various algorithms like LSTM, lasso and regression trees.

Agriculture deals with large datasets and knowledge process[12]. Many techniques are used to predict the yield of the crops. Neural pathways and sophisticated systems are used for massive volumes of data[18]. They used a replacement algorithm to determine yield in this research study. The pH level, amount of nitrogen, amount of carbon, weather patterns, temp, types of soil, and amount of phosphate are among the parameters. The primary goal of agriculture is to maximise value. Because large-scale climatological phenomena have an overly negative impact on agriculture, it is critical to clarify the rainfall patterns of a selected point. There is a lot of research going on with pulses, wheat, rice, sugarcane, and onions. More accurate predictions using meteorological data is an intelligent system. Statistical models and crop simulation models are the two groups of prediction models and applications[14]. ANN and genetic algorithms are efficient than traditional methods because there are easy and accurate for complex inputs. If ANN uses climate factor effective then farmers can use it efficiently.

ANNs are computational systems that processes like biological networks that constitute animal brains. Learning through experiences known as computational ability. Classification is made up of a group of simple computing units known as processing elements that are associated together in a multifaceted communication system, much like the human brain. Feed forward and back propagation are two common neural network architectures. A feed forward system as well as interpretation is a network that has no feedback mechanism. The flow of the system is uni-directional. A node transmits messages to another access point but obtains no response.

Fig.4. Back propagation

Back propagation seems to be a supervised classifier learning that describes how a neural network is trained. Network should provide with sample inputs and desired outputs. Following computations, the final output is compared to the actual output for a given input. Back propagation has
three phases: feed forward, back propagation of error, and weight updating.

Estimation of wheat crop yield have used a deep LSTM network [18] is a low-cost method of crop yield prediction because satellite images are obtained from publicly available sources. Tehsil (block) levels are proposed for wheat predictions across several states in India and outperformed the existing methods. Remote sensing data is important popular source of data for various applications like income prediction, yield prediction etc.while prior methods involve in extracting handcrafted or rudimentary features such as histograms where as this paper works on satellite images and allows model for learning yield prediction. For geographical area yield estimates depends on factors like waterbodies, urbanization etc. For data temporal features are modelled using deep LSTM model. Evaluation and validation of approach is done for tehsil level wheat prediction for seven states in India. To train proposed deep learning models used MODIS surface reluctance multispectral satellite images and land classification maps. The model outperformed traditional remote sensing methods by 70% and deep learning models by 54%.

Crop yield prediction using deep Gaussian process based on remote sensing data [15] says that it is an inexpensive method for accurate prediction. In three ways existing techniques are improved firstly they forego traditional features and next they undergo modern approaches based on remote sensing. Novel dimensionality reduction technique is introduced to train CNN or LSTM network based on labeled training data it will automatically learn the features. Finally, in order to improve accuracy, spatiotemporal data, the Gaussian model is expressly utilized. We put our technique to the assessment on province soybean predictive modeling in the U.s, and it surpasses competing firm procedures.

A new dimensionality reduction technique is used to fulfill the demands of training phase. Raw images are treated as histograms of pixel counts to achieve tractability mean field[15] approximation is used. CNN and LSTM are trained on histograms to predict the yield. It does not explicitly account for spatiotemporal dependencies between data points but performs well, for example, due to common soil properties. we have a tendency to overcome this limitation by incorporating a mathematician method layer on prime of our neural network[10] models. Other traditional remote sensing-based models were outperformed by 30 percent in aspects of RMSE and 15 percent of its total in aspects of MAPE. Deep learning models are complex non linear mappings used for learning hierarchical representation of data. CNN, DNN, LSTM consists of the set of layers where the output of one layer is the input of the next layer.
4. Methods For Predicting Crop Yield

4.1 Artificial Neural Networks (ANNs)

Artificial neural networks performed better than multivariate regression. Back propagation neural network technique is one of the best performing algorithm. Artificial neural network uses back propagation. The multilayer perceptron seems to be the most frequently used neural network in latest studies. The input data, the hidden neurons, and the network output are required for ANN, and it outdoes prediction models[2]. Biological neural mechanisms in the human brain are one type of concept for ANN structure. Neurons in the human brain are integrated and used to design the interdependencies for refining. A neurological network is made up of nodes or units, which are integrated structures. Every component is built to look like its biological counterpart, a neuron. Each unit receives and replies to a weighted set of input data. The Artificial Neural Network (ANN) [19] technique is based on a biological system model. The crucial element of this technique is the novel structure of the understanding process system. Associate degree ANN is developed for a single application, including such pattern classification or information classification, using a learning method.

4.2 Deep Learning Networks (DLNs)

Deep learning models are neurons that are used in deep learning approaches. CNN, RNN, LSTM, and DNN are outscored by Lasso, shallow neural networks, regression models, deep learning methods, fully convolutional neural networks, and rnns. Deep neural networks stand to gain from cutting-edge simulation and methods applied. Feature selection can be accomplished by decreasing the dimension of the input space used to train the DNN model. As it has numerous non-linear stacked layers which convert the input to the system into an elevated and more conceptual representation to every stacked layer, it can find the underpinning representation of data.
in the apparent lack of feature input data. More complicated features are extracted as the system grows deeper, resulting in high precision. It is known to be an ubiquitous approximator function if the right variables are provided, which means it can estimate almost any feature but is difficult to choose the right parameters.

4.2.1 Shallow Neural Network

One or multiple hidden layers are prevalent in shallow neural network models. Recognizing a shallow neural network can make you realize what tends to happen in a deep network. The section outlines a shallow neural network with one hidden units, one input nodes, and one output neurons.

![Shallow Neural Network](https://example.com/shallow_neural_network.png)

**Fig.10.** Shallow Neural Network

4.2.2 Convolution Neural Network

It is indeed a deep learning method that generates an image representation and designates weight values to various objects in the scene to discern them from each other. The below figure is the convolution neural network is sequence for classifying handwritten digits.

![Convolution Neural Network](https://example.com/convolution_neural_network.png)

**Fig.11.** Convolution Neural Network

4.2.3 Recurrent Neural Network (RNN)

The recurrent neural network model [20] adds a twist for basic neural network. A vanilla neural network takes input as fixed vector size and limits the usage to situations which involves series type input with no planned size. Recurrent neural network remembers the past and are influenced by the decisions in past. They can take one or more inputs and gives one or more outputs as vector.

![Recurrent Neural Network](https://example.com/recurrent_neural_network.png)

**Fig.12.** Recurrent Neural Network
4.2.4 Long Short Term Memory (LSTM)

Long Short Term Memory Models (LSTMMs) are designed essential to tackle the long-term correlation issue that exists in recurrent neural networks due to the vanishing gradient troubles. To make more traditional feed forward neural network LSTMs have feed back connections. They process the sequence of data entirely without treating each point independently and the useful information is retained about the previous data and it helps for processing the new data points. For text, speech and general-time series LSTMs are good for such sequences of data.

4.2.5 Deep Neural Networks

For association of input and outputs deep learning algorithms are used along with networks. Deep refers to large amount of layers along with weights and biases can be able to solve for more complex functions.

![Fig.13. Deep Neural Networks](https://doi.org/10.1051/e3sconf/202130901031)

4.3 Machine Learning Techniques

Machine learning models used to anticipate crop growth include regression analysis, gradient boosting regressor model, random forest regressor model, decision tree regressor mode, polynomial regression technique, and ridge regression technique. Out of all random forest regressor[3] and gradient boosting regressor[3] gives best accuracy with cross validation.

4.3.1 Linear Regression Model

It is just a sequential model that represents the linear relation among an input parameter (x) and a single output unit (y) for combinations of the input parameter (x) (y). Single input variable is considered as a simple linear regression and multiple input variable is considered as a multiple linear regression. Most common data and method is treat the ordinary least squares method.

4.3.2 Gradient Boosting Regressor

Gradient Boosting is a kind of regression method. GB builds an ingredient model series of steps, letting the tuning of absolute discrete loss functions. A regression tree performs the amount of boosting phases in each phase by working on the negative slope of the given deficit.

4.3.3 Random Forest Regressor

So every decision tree has a large variance; nevertheless, when we incorporate them all at once, the eventual variance is lesser because each tree is wholly trained on suitable sample knowledge, and thus the outcome does not rely solely tree structure but on multiple call trees. In the occurrence of a classification error, the bulk vote classifier is used to produce the final output. In the scenario of a regression flaw, the final output is the mean of all the outcomes. Aggregation is the name given to this part of the equation.

4.3.4 Decision Tree Regressor

In attempt to provide tangible continuous output, decision tree predictor scrutinises an object's choices and trains a framework within the structure of a tree to anticipate knowledge in the future. Constant output insinuates that the emission is not distinguishable, that it cannot be characterised simply by a distinct, well-known group of statistics or value systems.

4.3.5 Polynomial Regression

Polynomial Regression is a common formula that does use an ordinal degree polynomial to quantify the relationship between a predicated (y) and an independent variable (x). The Unique Instance of Multiple Statistical Residuals in Cubic Cm is another name for it. As a result, in required to persuade the Multiple Statistical Coefficient of determination to Polynomial Correlation, we add some polynomial conditions to it. It is a linear classifier that has been improved in high precision. The mentoring set of data used in Polynomial regression is not deterministic. It hires a statistical regression model to adapt the tough and non-linear features and data sources.

4.3.6 Ridge Regression

Ridge regression is a concept standard setting technique that might be used to analyse any statistics that has experienced from repeated measures. This method employs L2 regularisation. Once the problem of multiple regression happens, least-squares ar unbiased, and variances ar giant, this ends up in foreseen values to be far-flung from the particular values.
5. Comparison of crop yield prediction with different models

5.1 Artificial Neural Networks

The table below contrasts R, MEA, and MSE[16] for crop varieties when using artificial neural systems imperialistic competitive optimization technique and grey wolf optimization method models to examine effectiveness.

<table>
<thead>
<tr>
<th>Crops</th>
<th>R</th>
<th>MEA (%)</th>
<th>RMSE (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ANN-ICA</td>
<td>ANN-GWO</td>
<td>ANN-ICA</td>
</tr>
<tr>
<td>Wheat</td>
<td>0.35</td>
<td>0.49</td>
<td>35.6</td>
</tr>
<tr>
<td>Barley</td>
<td>0.38</td>
<td>0.43</td>
<td>12.10</td>
</tr>
<tr>
<td>Potato</td>
<td>0.81</td>
<td>0.81</td>
<td>22.8</td>
</tr>
<tr>
<td>Sugar Beet</td>
<td>0.24</td>
<td>0.26</td>
<td>39.0</td>
</tr>
<tr>
<td>Average</td>
<td>0.45</td>
<td>0.50</td>
<td>27.3</td>
</tr>
</tbody>
</table>

5.2 Deep Neural Networks

The table compares different deep learning models based on response variable, training RMSE, validation RMSE, training and validation correlation coefficients for yield, check yield, and difference.

<table>
<thead>
<tr>
<th>Response</th>
<th>Model</th>
<th>Training RMSE</th>
<th>Training correlation coefficient(%)</th>
<th>Validation RMSE</th>
<th>Validation correlation coefficient(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Corn</td>
<td>CNN-RNN (W)</td>
<td>17.15</td>
<td>89.02</td>
<td>24.79</td>
<td>72.18</td>
</tr>
<tr>
<td></td>
<td>CNN-RNN (S)</td>
<td>19.16</td>
<td>84.72</td>
<td>24.43</td>
<td>73.82</td>
</tr>
<tr>
<td></td>
<td>CNN-RNN (M)</td>
<td>27.77</td>
<td>69.59</td>
<td>34.03</td>
<td>33.99</td>
</tr>
<tr>
<td>Average</td>
<td></td>
<td>38.51</td>
<td>0.04</td>
<td>35.66</td>
<td>0.01</td>
</tr>
</tbody>
</table>

| Soya bean | CNN-RNN (W) | 4.6 | 89.65 | 5.98 | 78.80 |
|           | CNN-RNN (S) | 5.64 | 83.26 | 5.94 | 81.28 |
|           | CNN-RNN (M) | 7.87 | 59.84 | 8.66 | 48.77 |
| Average   |       | 10.29 | 0.01 | 10.11 | 0.05 |

5.3 Machine Learning

When accuracy, recall, and F1-score are being used to make comparisons different algorithms such as logistic regression model, decision tree method, Random Forest model, and k nearest neighbour as well as svms, logistic regression model performed the best.
Table 4. Machine learning algorithms accuracy comparison

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>100</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.90</td>
<td>0.83</td>
<td>1</td>
<td>93.3</td>
</tr>
<tr>
<td>Random Forest</td>
<td>0.90</td>
<td>0.83</td>
<td>1</td>
<td>93.3</td>
</tr>
<tr>
<td>K Nearest Neighbor</td>
<td>0.82</td>
<td>0.63</td>
<td>0.8</td>
<td>86.6</td>
</tr>
<tr>
<td>Support Vector Machine</td>
<td>0.60</td>
<td>0</td>
<td>0.75</td>
<td>60</td>
</tr>
</tbody>
</table>

6. Conclusion

This paper gives a detailed analysis of different approaches used for crop yield prediction. Various algorithms are mentioned along with the datasets and results. Different types of crops used for prediction purpose in different seasons are mentioned. Various crops are used for prediction using different classification techniques, depending on the factors used more accurate results are generated. Climatic conditions play a major role in yield. Crop yield prediction mainly depends on factors so while using any techniques need to consider the factors accurately gives effective results and helps in making good decisions for import and export and better pricing of respective crops. This work can be used to extend these approaches for and to achieve better results for future enhancements.

7. Nomenclature

CNN - Convolutional Neural Network
RNN - Recurrent Neural Network
RT - Regression tree
DL - Deep Learning
ANN - Artificial Neural Networks
GB - Gradient Boosting
DRL - Deep Reinforcement Learning
ARIMA - AutoRegressive Integrated Moving Average
LSTM - Long Short Term Memory
DNN - Deep Neural networks
MAPE - Mean Absolute Percentage Error
RMSE - Root mean square error
MEA - Means end analysis
MSE - Mean square error
BDN - Deep Belief Network
BAN - Born again neural networks
IDANN - Internal Deep Generative Artificial Neural Networks
RAE - Robust Adaboost RT based Ensemble
R - Range
MODIS - Moderate Resolution Imaging Spectroradiometer
LASSO - Least Absolute Shrinkage and Selection Operator
ANN-ICA - Artificial Neural Networks - imperialist competitive algorithm
ANN-GWO - Artificial Neural Networks -Gray Wolf Optimizer of Neural Networks
CNN-RNN(W) - Convolutional Neural Networks-Recurrent Neural Networks( weather data)
CNN-RNN(S) - Convolutional Neural Networks-Recurrent Neural Networks( soil data)
CNN-RNN(M) - Convolutional Neural Networks-Recurrent Neural Networks( management data)
8. References