Original Research Article
Modeling cereal yield forecasting using Machine learning techniques for Shri Ganganagar, Rajasthan, India

ABSTRACT

One of agriculture's most difficult issues is predicting crop yield. Crop yield forecasting enables necessary decisions to be made to guarantee food security. The current study looks into the use of statistical and machine learning approaches for wheat and rice yield prediction, using long-term weather and yield data of ShriGanganagar, Rajasthan, India. Weather-based models may give accurate crop production estimates, but choosing the right model for agricultural output projections can be difficult. As a result, different models were compared in this study to determine the best model for rice and wheat yield prediction, including Multiple linear regression (SMLR), an Aartificial neural network (ANN), the Least absolute shrinkage and selection operator (LASSO), an Eelastic net (ELNET), Ridge regression, and K-nearest neighbor (K-NN). K-NN outperformed ANN in rice crops and fared better in wheat crops based on the lowest nRMSE value. Ridge regression based on nRMSE was the next best model for Rice yield prediction in the examined region, and KNN was the second best model for Wheat crop.

Keywords: Lasso, Ridge, Machine learning, ANN, KNN, PCA, Yield forecasting

1. INTRODUCTION

Agriculture, together with its linked sectors, is without a doubt the most important source of income in India, particularly in the vast rural areas. It also makes a substantial contribution to the Gross Domestic Product (GDP). In 1950, 55% of Gross Domestic Product (GDP) came from agriculture while in 2009 it is 18.5% and during the financial year 2015-2016 it is 16.85% **(1).** According to the report in 2018 near about 15.4% of GDP and 50% of manpower is contributed by agr and its allied activities **(2).** Green revolution has ushered heights in Indian agriculture ensuring food security to its burg population with 272 million tons of food grain production in 2017-18. Tentative requirement in 2030 and 2050 is 3 494 million tons respectively. The human population of India increased to 1.21 billion in 2011 at a growth rate of 1.71% over 1.03 billion in 2001 and expected to be 1.53 billion in 2030. The country will face a tremendous worth to feed the population and meet the nutritional security to 17.5% of the stuff population and fed 11% of the world's livestock population working smoothly on 2.3% of land and fresh water resource of 4%. The per capita shrinkage of cultivable land from 0.34 hectare in 1950-51 to 0.08 hectare in 2025 has been a matter of concern in India **(3)** A huge portion of fertile cultivable land has been grabbed for urbanization, industrialization and other works for which, since 1970 the cultivated area has been lowering about 141 ± 2 million hectare **(4).**Sustainable agriculture is critical for comprehensive rural development in terms of food security, rural employment, and ecologically friendly technologies such as soil conservation, sustainable natural resource management, and biodiversity protection.

Climate change and its aftereffects are among the new issues of the period. The rise in the universal mild outside air temperature is attributed to an impending climate change. The combined effects of expanding population, natural weather variability, soil loss, and climate-changing require procedures to assure timely and consistent agricultural growth and output. It is also necessary to contribute to the expansion of agricultural food production sustainably. According to the FAO (Food and Agriculture Organisation), demand for and consumption of grains has increased dramatically in comparison to output in emerging countries such as India. Demand for rice, wheat, and other coarse grains will have increased between 1964

Commented [A1]: The manuscript may be written in past tense
The share of agriculture in GDP may be given
for the year 2023-24, 2015-16 is too old data

and 2030. To accommodate expanding demand, developing-country cereal imports expanded dramatically, rising from 39 million tonnes per year in 1970 to 130 million tonnes per year by 1997-1999. Import growth is expected to continue and may accelerate in the future years. These emerging countries are expected to import 265 million tonnes of grains by 2030, accounting for 14% of their total annual consumption. Nations that do not consider taking measures to reduce their total reliance on imports for conventional crops may suffer substantially as a result of these situations. As a result, it is a worldwide challenge to change the existing situation in the future and make nations more self-sufficient in fulfilling their food demands, which necessitates accurate and timely crop yield predictions. Crop production prediction is one of the most challenging jobs in agriculture. In many countries, losses due to weather conditions account for up to 30% of the annual agriculture production (**5**). Therefore, there is a high demand to develop models that give accurate yield prediction before harvest which can be used by the government, policymakers and farmers for making advance planning and strategies.

Nonetheless, in recent years, the growth of new technology, such as crop model simulation and machine learning, has shown to estimate yield more precisely, as well as the ability to analyse massive amounts of data using high-performance computers. In the current scenario, forecasting of crop yield using Artificial Neural Network (ANN), Least Absolute Shrinkage and Selection Operator (LASSO) and Elastic Net (ELNET) getting a great deal of attention (**6, 9, 23**). Various efforts have been made by the researchers to develop pre harvest yield forecast models based on yield and weather-datasets. Numerous investigations are presently exhibiting relatively higher potential in the use of machine learning algorithms than traditional statistics (**9, 10, 11**). Machine learning is an area of artificial intelligence where computers can be taught without certain programming. Such approaches overcome agricultural structures, which are both non-linear or linear, by ensuring a notable prediction capacity (**14**). The strategies are obtained from the learning method in the machine learning agricultural system. These methods involve on carrying out a particular task through the train with the training information.

Commented [A2]: what does it mean? Is it correct?

Unfortunately, little scientific effort has been undertaken till date to construct a yield forecast model for the Rajasthan area using machine learning approaches. So far, the majority of research has relied on predictions based on standard statistical models. As a result, in the current study, an attempt was made to provide a yield forecast of the rice and wheat crop for the ShriGanganagar district of Rajasthan using SMLR, ANN, LASSO, ELNET, and Ridge regression, and a comparison was made among these techniques to select the best model that can be used to provide rice and wheat crop yield forecasts for the districts.

2. MATERIAL AND METHODS

Time series data of Rice and Wheat crop of twenty years in Sri Ganganagar district for the period 2001 – 2021 has been taken from DACNET district level yield database. The changes in yield time series data over the years are impacted by technological differences, environmental variability, and other factors, resulting in nonlinear and non-stationary trends that must be eliminated before computing the fundamental correlation function to improve the model's prediction performance (**15)**. De-trending the yield data is thus required to eliminate the longterm mean changes from time series.One of the most common techniques is to run a preset function versus time, such as a basic linear regression model or a secondorder polynomial regression model. Many researchers have used this approach to de-trend agricultural yield data and investigate the effects of climatic variability (**16, 18, 29**). In the current study, a simple linear regression model was used to detrend rice and wheat production. Simple linear regression model can be fitted against time using the method of least squares.

$$Y_T = \beta_0 + \beta_{1T} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (1)$$

Where, Y_T is the crop yield at time t; time t is the predictor, and β_0 and β_1 are the coefficients. The residuals (detrended yield) of this model were used for indices calculation.

Commented [A3]: What is $\beta_{1T?}$
There shuold be error term in eq 1

Daily data of maximum & minimum Temperature (°C), Relative humidity morning (%), Relative humidity evening (%), Sunshine hours (hours/day) and Rainfall (mm) for the period 2001-2021 were collected from the KVK Sri Ganganagar, Rajasthan. PET data was calculated for the same time period using FAO-CROPWAT software. For the rice crop 26th to 49th standard meteorological week (SMW) weather data and for wheat crop 43rd to 13th week weather data were considered in the study to develop the yield forecast model. Weekly average was calculated from the daily weather data. These average values are then used for the calculation of weighted and unweighted weather indices.

Commented [A4]: Is it defined earlier?

Unweighted indices were generated by summation of individual or interaction of weather variables, whereas weighted indices were generated by summation of individual or interaction of weather variables and its correlation with detrended yield of Rice and Wheat. The first index (Unweighted) represents the total amount of weather parameters received by the crop during the period under consideration, whereas the latter (weighted) represents distribution of weather parameters (**20**). The formula for the calculation of Unweighted and weighted weather indices are mentioned below.

Table 1.Unweighted and weighted weather indices for the development of multivariate models.

Mean weekly weather variable	Unweighted Weather indices	Weighted Weather indices
Tmax (Tx)	Z10	Z11
Tmin (Tn)	Z20	Z21
Rainfall (R)	Z30	Z31
Solar Radiation (RAD)	Z40	Z41
Relative Humidity I	Z50	Z51
Relative Humidity II	Z60	Z61
PET	Z70	Z71
Tmax*Tmin	Z120	Z121
Tmax*Rainfall	Z130	Z131
Tmax*Solar Radiation	Z140	Z141
Tmax* Relative Humidity I	Z150	Z151
Tmax*Relative Humidity II	Z160	Z161
Tmax*PET	Z170	Z171
Tmin*Rainfall	Z230	Z231
Tmin*Solar Radiation	Z240	Z241
Tmin* Relative Humidity I	Z250	Z251
Tmin*Relative Humidity II	Z260	Z261
Tmin*PET	Z270	Z271
Rainfall*Solar Radiation	Z340	Z341
Rainfall* Relative Humidity I	Z350	Z351
Rainfall*Relative Humidity II	Z360	Z361
Rainfall*PET	Z370	Z371
Solar Radiation * Relative Humidity I	Z450	Z451
Solar Radiation *Relative Humidity II	Z460	Z461
Solar Radiation*PET	Z470	Z471
Relative Humidity I *Relative Humidity II	Z560	Z561
Relative Humidity I*PET	Z570	Z571

Relative Humidity II*PET	Z670	Z671
Tx = Maximum Temperature (°C), Tn = Minimum Temperature(°C), R = Rainfall (mm), RAD = Bright Sunshine hours (hours), Relative Humidity I = Relative humidity during morning hours (7:20 am), Relative Humidity II = Relative humidity during afternoon hours (2:20 am), Unweighted weather indices = simple total of values of weekly weather variables in different weeks, Weighted weather indices = weighted total of mean weekly values, where correlation coefficients have been used to compute weights.		

.

2.1 UNWEIGHTED WEATHER INDICES:

$$Z_{ij=} \sum_{w=1}^{m} X_{iw} \ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

$$Z_{ii'j} = \sum_{w=1}^{m} X_{iw} X_{i'w} \ldots\ldots\ldots\ldots\ldots \quad (3)$$

2.2 WEIGHTED WEATHER INDICES:

$$Z_{ij} = \sum_{w=1}^{m} r_{iw}^{j} X_{iw} \ldots\ldots\ldots\ldots \quad (4)$$

Commented [A5]: Use only three dots.

$$Z_{ii"m} = \sum_{w=1}^{m} r_{ii'm}^{j} X_{iw} X_{i'w} \ldots\ldots\ldots\ldots\ldots \quad (5)$$

$X_{iw}/X_{ii'}$ = value of *i* th/*i'*th weather variable under study in wth week,

$r_{iw}^{j}/r_{ii'}^{j}$ = correlation coefficient of yield with ith weather variable/product of *i*th and *i'*th weather variables in wth week

M = week of forecast

In the present study, to develop a good crop yield prediction model, eight differenttypes of multivariate analysis techniques are used. Details of those models are given as follows:

Commented [A6]: Multivariate or modeling techniques.

2.3 Linear Regression using ML

Linear Regression is a statistical modelling technique used to establish correlations between independent and one or more dependent variables. In machine learning, linear regression utilises data to learn by minimising loss (usually referred to as RMSE or MSE) using methods such as gradient descent. According to the nature of the data, the gradient descent technique fits the models at minimised loss functions, which increases the predicted accuracy of the model. Usually, Linear Regression is defined by the following equation:

$$Y_i = f(X_i, \beta) + e_i \ldots\ldots\ldots\ldots \quad (6)$$

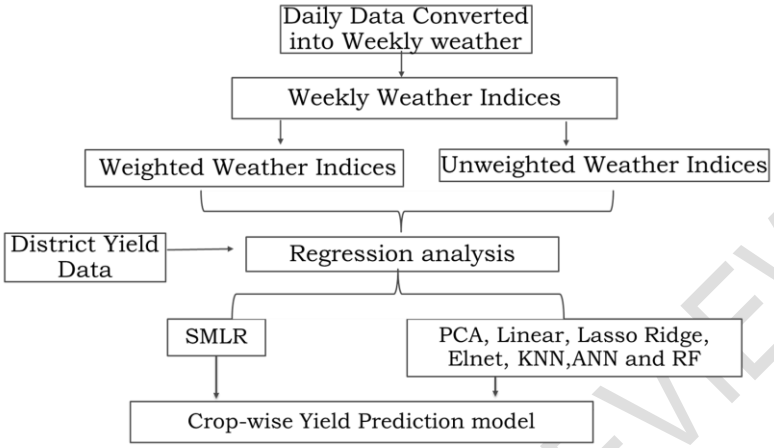

Fig.1 Flowchart of Methodology

Commented [A7]: Discussion on fig 1 was not found in text

2.4. Principal Component Analysis- linear regression

In our study, we used principal component analysis (PCA) on all-weather indices for each district of both crops. PCA seeks to decrease the dimensionality of a data collection while maintaining the majority of the information. The combination of feature extraction and selection approach for data analysis ~~is~~ was PCA followed by linear regression to develop yield prediction model. PCA scores were used as input for MLR analysis. To solve the problem of multicollinearity among weather variables, PC scores were employed as regressors for SMLR and ANN in the development of agricultural yield models (**19**).

2.5. Artificial Neural Network using ML

In this study, we employed three layers of artificial neural networks (ANN): input, hidden, and output feed-forward. Each layer is made up of neurons or nodes that are linked to one another. The dataset determines the number of nodes in the input and output layers.The input layer consists a number of neurons which are equivalent to the number of input feature and the output layer has only one neuron which is crop yield. The number of hidden neurons may vary based on the number of input features. A neural network consists of a set of highly inter-connected entities, called nodes or units. ANN can derive relationship between input and output on any process.

2.6. Least absolute shrinkage and selection operator (LASSO)

The least absolute shrinkage and selection operator (LASSO) is a regression technique that does both variable selection and regularisation to improve the statistical model's prediction accuracy and interpretability. By using LASSO, which penalises the coefficients of the regression variables and reduces some of the coefficients to zero, this strategy helps to prevent over fitting. As a result, the remaining input variables with non-zero coefficients following the shrinking process are chosen in this phase to be included in the model. The purpose of LASSO is to minimize the forecasting error **(25).** LASSO regression uses L1 regularization technique. The objective function that is minimized by the LASSO algorithm is expressed as

Commented [A8]: What is L1 regularization technique. May be discused in brief.

$$L_{lasso}(\widehat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\widehat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}| \ldots\ldots\ldots\ldots\ldots\ldots \quad (7)$$

2.7 Ridge regression

Ridge regression is a strategy for decreasing data over fitting by biasing regression results somewhat. The main purpose of employing ridge regression is to achieve more accurate results. When there is a strong correlation between the predictor variables, the approach allows for the estimate of coefficients in multiple regression models **(26).** Ridge regression may perform slightly poorly on the training set, but overall, it performs consistently well. It uses L2 regularization technique. The loss in ridge regression is defined as:

$L_{ridge}(\widehat{\beta}) = \sum_{i=1}^{n}(y_i - x_i^j \widehat{\beta})^2 + \lambda \sum_{j=1}^{m} \beta_j^2$…… (**7**)

2.8 Elastic net (ENET)

ELNET regression uses both L1and L2 regularization techniques of LASSO and ridge regression to improve model performance **(27).**L1 regularisation, also known as lasso regression, adds the coefficient's "absolute value of magnitude" as a penalty term to the loss function. L2 regularisation, also known as ridge regression, adds the "squared magnitude" of the coefficient to the loss function as a penalty term. For ELNET regression the loss is defined as

$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda\left(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}^2 + \alpha\sum_{j=1}^{m}|\widehat{\beta}_j|\right)$……..(**8**)

2.9 K-Nearest Neighbor

The (K-NN) k-nearest neighbor is used for both classification and regression problems. It is one of the simplest classification algorithms based on Supervised Learning technique. It works by determination of the parameter ***k*** which is number of nearest neighbors. When there is new data point to classify, then its k-nearest neighbors is find out from the training data by calculating the distance between the input variable and the all the data points in the dataset. It use proximity to make classification and predictions about the grouping of an individual data point. Regression problems use the average the k nearest neighbors is taken to make a prediction about a classification. The larger is k; the better is classification **(28, 30).**

2.10 Random Forest

Random Forest is a supervised learning algorithm. It creates a forest and makes it somehow random. Random Forest is a flexible, easy to use machine learning algorithm that produces, even without hyper-parameter tuning, a great result most of the time. It is also one of the most used algorithms, because its simplicity and the fact that it can be used for both classification and regression tasks which form the majority of current machine learning systems **(31).**

2.11 Model performance

The performances of different models were compared using Root Mean Square Error (RMSE), Normalized root mean square error (NRMSE) and mean square error (MSE). The lower the value of these error, the better it is.

Commented [A9]: What is n RMSE in abstract? Both are same or different

3. RESULTS AND DISCUSSION

3.1 Evaluation of the Model Performance

The results pertaining to all models' performances and equations developed using these models are shown in Tables 2 and 3 and 4, respectively. It may be observed from the models developed for both the crops that weighted weather variables have greater frequency than un weighted variables.

3.1.2 Linear regression

The performance of Linear regression was good at the calibration stage for both the crops, with a R^2value of 1.0 for Rice and Wheat crop with a RMSE of 0.0 and 0.0, kg/ha respectively. The nRMSE values for the calibration stage were also zero indicating good model performance, though at the validation stage, the performance of linear regression was poor for both the crops, with R^2 values of 0.43 for rice crop and 0.41 for wheat crop. In addition to this, the nRMSE value at the validation stage indicated good model performance for Rice crop (nRMSE = 4.34%) and fair performance for Wheat crop (nRMSE = 14.11%). The RMSE value indicates that the model overestimated the crop yield for the rice crop (RMSE = 123.63kg/ha) and wheat crop (RMSE = 504.19kg/ha) at the validation stage. The RMSE value also suggested the better model performance for rice crop as compared to the wheat crop.

Commented [A10]: What was number of data points? Because R2
Increase when we increase no of independent variable. Here R2 is one that is no error and model was able to capture
100% variablity in data but it failed for new observation

Table 2. Models developed using multivariate regression techniques for Rice crop.

Commented [A11]: ANOVA table may also be presented

TECHNIQUES	***EQUATIONS***

Commented [A12]:

LINEAR	Yield= -16388 - 0.05316 x Z10 + 0.003954 x Z11 - 0.00742 x Z20 - 0.00129 x Z21 - 0.00964 x Z30 - 0.00159 x Z31 + 0.006824 x Z40 + 0.007717 x Z41 - 0.0317 x Z50 - 0.01325 x Z51 + 0.008566 x Z60 + 0.007389 x ZZ61 - 0.01849 x Z70 + 0.00602 x Z71 + 0.184047 x Z120
LASSO	Yield = 4253.1 – 33.2158 x Z10 + 4.471016 x Z20 + 0.001763 x Z30 + 7.95687 x Z31 + 10.43094 x Z40 + 42.6914 x 6Z41 + 0.22427 x Z50 - 5.15707 x Z51 + 2.4561 x Z60 - 1.11145 x Z61 - 4.45852 x Z70 + 0.32893 x Z120
RIDGE	Yield = -12071 - 0.004745 x Z10 - 0.000955 x Z11 - 0.00364 x Z20 - 0.002456 x Z21 - 0.024436 x Z30 + 0.000795 x Z31 + 0.005296 x Z40 + 0.005134 x Z41 - 0.023634 x Z50 - 0.011285 x Z51 + 0.01129 x Z60 + 0.005982 x Z61 - 0.016147 x Z70 + 0.004076 x Z71 + 0.245439 x Z120
ELNET	Yield = -5723 - 9.08724 x Z10 - 0.68167 x Z20 + 0.916937 x Z30 + 5.438448 x Z31 - 1.17095 x Z40 + 0.268121 x Z41 + 2.148623 x Z50 - 1.8131 x Z51 + 2.486351 x Z60 - 1.57291 x Z61 - 10.7841 x Z70 + 0.172515 x Z71 + 0.245555 x Z120

Table 3.Models developed using multivariate techniques for Wheat crop.

TECHNIQUES	***EQUATION***
Linear	Yield = 7519.6+0.183699 x Time+0.045227 x Z10 +0.017922 x Z11–0.02183 x Z20 +0.041842 x Z21 – 0.00797 x Z30+0.009618 x Z31 +0.01629 x Z40 – 0.02247 x Z41 – 0.37357 x Z50+0.205239 x Z51 +0.268943 x Z60 –0.07749 x Z61 –0.00662 x Z70 – 0.01565 x Z71
LASSO	Yield = -199.6+119.16 x Time-0.62988 x Z10+0 x Z11+0 x Z20 +3.602692 x Z21–0.46358 x Z30+0 x Z31+10.30657 x Z40+0 x Z41+2.241803 x Z50+6.697307 x Z51+1.879091 x Z60+3.481065 x Z61 +0.467174 x Z70+0 x Z71
RIDGE	Yield = 7519.6+0.183699 x Time+0.045227 x Z10 +0.017922 x Z11–0.02183 x Z20 +0.041842 x Z21 –0.00797 x Z30+0.009618 x Z31 +0.01629 x Z40 –0.02247 x Z41 – 0.37357 x Z50+0.205239 x Z51 +0.268943 x Z60 –0.07749 x Z61 –0.00662 x Z70 – 0.01565 x Z71
ELNET	Yield = 12963+23.99188 x Time-14.3463 x Z10+0 x Z11+10.98483 x Z20-0.59521 x Z21 –0.00952 x Z30-0.76866 x Z31- 5.58891 x Z40 –0.47459 x Z41 –4.26642 x Z50+18.2113 x Z51 +3.587855 x Z60 –10.0836 x Z61 –7.06053 x Z70 –0.0328 x Z71

Where: Z10 = Maximum Temperature (°C), Z20 = Minimum Temperature (°C), Z30 = Rainfall (mm), Z40 = Solar Radiation (hrs.), Z50 = Relative Humidity I (%), Z60 = Relative Humidity II (%), Z70 = PET, Z120 = Tmax*Tmin. Z11 = Maximum Temperature (°C), Z21 = Minimum Temperature (°C), Z31 = Rainfall (mm), Z41 = Solar Radiation (hrs.), Z51 = Relative Humidity I (%), Z61= Relative Humidity II (%), Z71 = PET

3.1.2 Ridge

The results of the Ridge analysis revealed that the value of R^2 ranged from 1.0. The highest R^2 was observed for the Rice crop (R^2= 1.0), with a RMSE value of 0.0kg/ha, followed by Wheat crop (R^2= 1.0) with RMSE values of 0.01kg/ha

Commented [A13]: What was value of ridge hper parameter for this?

respectively. On the other hand, in the validation stage, the R^2 value ranged from 0.41 to 0.83. The highest R^2 was observed for Rice crop (R^2= 0.83), with RMSE 84.98kg/ha, whereas the lowest R^2 was observed for Wheat crop (R^2= 0.41), with RMSE 504.18kg/ha. Moreover, at the validation stage, the performance of the Ridge model was good for Rice crop (nRMSE = 2.98%) and fair for wheat crop (nRMSE = 14.11%). Hence, the Ridge model can be used to forecast the rice yield for the Shri Ganganagar district.

3.1.3 Least Absolute Shrinkage and Selection Operator (LASSO)

For LASSO regression, at the calibration stage, the R^2 value was found for the Rice crop (0.99), with a RMSE of 1.05kg/ha, and R^2 was recorded for Wheat crop (0.99), with a RMSE of 1.18 kg/ha. In addition to this, the value of the nRMSE showed good model performance (nRMSE< 0.04%) for the both crop. The RMSE statistic for the training showed that the performance of the LASSO regression model was good for Rice crop (109.12kg/ha) and fair performance for Wheat crop (430.19kg/ha). At the validation stage, the performance of the LASSO model was good for Rice crop (nRMSE = 3.83%) and fair for wheat crop (nRMSE = 12.04%).Contrary to this, at the testing stage, the value of R^2 was good for Wheat crop (R^2 = 0.88) and fair for Rice crop (R^2= 0.82).

3.1.4 Elastic Net (ELNET)

In case of the ELNET model, the values of the value of R^2 were 0.99 for the Rice and Wheat crop, at the training stage. The RMSE of training data was 1.62kg /ha and 7.52 kg/ha for rice and wheat crop respectively.In addition to this, the value of the nRMSE showed good model performance (nRMSE = 0.06%) for the Rice crop and fair Wheat crop (nRMSE = (0.21%). During the testing stage, the value of R^2 was good for Rice crop (R^2= 0.83, nRMSE = 4.09%) and Wheat crop (R^2= 0.75, nRMSE = 11.37%). The RMSE of testing data for Rice crop was 116.69kg/ha and Wheat crop (RMSE = 406.22kg/ha).

3.1.5 K-Nearest Neighbor (KNN)

The performance of K-Nearest Neighborwas good at the training stage for Rice crop with R^2 value of 0.47 and Wheat crop R^2 value 0.13 with a RMSE of 354.49kg/ha and 673.51kg/ha respectively. The nRMSE values for the training stage for Rice crop was 12.79% and Wheat crop was 18.55%. At the testing stage, the performance of K-Nearest Neighbor was good for the Rice crops, with R^2 values of 0.82 and 0.71 for Wheat crop. In addition to this, the nRMSE value at the testing stage indicated good model performance for Rice crop (nRMSE = 2.97%) and fair performance for Wheat crop (nRMSE = 4.67%). The RMSE value indicates that the model underestimated the crop yield for the rice crop (RMSE = 123.63kg/ha) and wheat crop (RMSE = 504.19kg/ha) during the testing stage. The nRMSE value also suggest better model performance for rice crop as compared to the Wheat crop.

Commented [A14]: What was k value (hyper parameter) used here? 3, 5,7etc

3.1.6 Random forest (RF)

For Random Forest, at the training stage, the R^2 value was found for the Rice crop (0.92), with a RMSE of 206.04kg/ha, and for Wheat crop (R^2 = 0.87), with a RMSE of 242.95kg/ha. In case of training the performance of wheat crop (nRMSE = 6.71%) was good based on nRMSE in comparison to rice crop with value of about 21.1%. The RMSE statistic for the testing stage showed that the performance of the Random forest model was good for Rice crop (108.30kg/ha) and fair performance for Wheat crop (281.18kg/ha). At the validation stage, the performance of the Random model was good for Rice crop (nRMSE = 4.03%) and for wheat crop (nRMSE = 7.58%).Contrary to this, at the testing stage, the value of R^2 was good for Wheat crop (R^2 = 0.81) and fair for Rice crop (R^2= 0.77).

3.1.7 Principal component analysis with linear regression (PCA-linear)

The results of the Principal component analysis exposed that during the training stage value of R^2 was observed for the rice crop to be 0.74, and for wheat crop (R^2 = 0.59), with a RMSE value of 206.04kg/ha for Rice crop followed by Wheat crop RMSE values of 359.56kg/ha respectively. The highest R^2 was observed for Wheat crop (R^2= 0.83), with RMSE 344.18kg/ha, whereas the R^2 was observed for Rice crop (R^2= 0.75), with RMSE 171.37kg/ha during testing. Moreover, at the testing stage, the performance of the PCA model was good for Rice crop with (nRMSE = 5.91%) and fair for wheat crop (nRMSE = 10.86%).

3.1.8 Artificial Neural Network (ANN)

The results of the analysis showed that the performance of the artificial neural network (ANN) was excellent for the Wheat crop, it was observed R^2= 0.71and for Rice crop R^2= 0.75 for the testing stage. A RMSE of both crop 260.48 and 290.24 kg/ha, for wheat and rice crop respectively, during validation. In addition to this, the value of the nRMSE for the Wheat and Rice crop was under 10% for the testing stage, the (nRMSE = 4.26%) for Wheat crop and (nRMSE = 6.90%) for Rice crop.

Commented [A15]: What was number of neurons in input, hidden
And output layer? What is learning rate? All these need
To be discused.

The study results revealed that during testing stage the order of performance of crop yield forecasting models developed for Rice crop based on nRMSE were as following: KNN (2.97) > Ridge (2.98) > Lasso (3.83) > RF(4.03) >Elnet (4.09) > Linear(4.34) >PCA_Linear (5.91) > ANN (6.90) > SMLR (13.57). Similarly the order of performance of yield forecasting models for Wheat crop, based on nRMSE was: ANN (4.26) > KNN (4.67) > RF (7.58) >SMLR (8.89) >PCA_Linear (10.86) >Elnet (11.37) >Lasso (12.04)> Ridge=Linear (14.11). This finding was in line with the study done by (**17**), which concluded that the performance of ANN was better as compared to SMLR, PCASMLR, LASSO and ELNET for Patiala district. It was found that ANN is a potential tool for developing in-season yield mapping and forecasting systems for corn in eastern Canada. They found that ANN yield models achieved better prediction accuracy (about 20% validation RMSE) than conventional models **(34)**. The another study (**33**) explored the application of statistical and machine learning methodologies for mustard yield prediction at eight sowing dates, utilising long-term (2006-2021) meteorological and disease data obtained from the experimental fields of GBPUA&T, Pantnagar, India. Cross-model comparisons revealed that ANN followed by LASSO can accurately forecast mustard yield at the majority of sowing dates.

Commented [A16]: The prediction may be presented for rice and wheat
Using all the used model.

Table 4. Comparison of different models for crop yield in training and testing datasets.

		Model	***LINEAR***	***RIDGE***	***LASSO***	***ELNET***	***KNN***	***RF***	***PCA***	***ANN***
Rice	Training	R^2	1.00	1.00	0.99	0.99	0.47	0.92	0.74	0.08
		MSE	0.00	0.00	1.10	2.61	125661.86	341225.33	42451.73	246897.21
		RMSE	0.00	0.00	1.05	1.62	354.49	584.14	206.04	496.89
		nRMSE	0.00	0.00	0.04	0.06	12.79	21.11	7.54	18.00
	Testing	**Model**	**LINEAR**	**RIDGE**	**LASSO**	**ELNET**	**KNN**	**RF**	**PCA**	**ANN**
		R^2	0.43	0.83	0.82	0.83	0.82	0.77	0.75	0.75
		MSE	15283.46	7222.15	11907.60	13616.40	6296.79	11729.53	29369.03	84237.27
		RMSE	123.63	84.98	109.12	116.69	79.35	108.30	171.37	290.24
		nRMSE	4.34	2.98	3.83	4.09	2.97	4.03	5.91	10.09
Wheat	Training	**Model**	**LINEAR**	**RIDGE**	**LASSO**	**ELNET**	**KNN**	**RF**	**PCA**	**ANN**
		R^2	1.00	1.00	0.99	0.99	0.13	0.87	0.59	0.13
		MSE	0.00	0.00	1.40	56.59	453622.03	59025.98	129283.90	591386.98
		RMSE	0.00	0.01	1.18	7.52	673.51	242.95	359.56	769.02
		nRMSE	0.00	0.00	0.03	0.21	18.55	6.71	9.69	22.20
	Testing	**Model**	**LINEAR**	**RIDGE**	**LASSO**	**ELNET**	**KNN**	**RF**	**PCA**	**ANN**
		R^2	0.41	0.41	0.88	0.75	0.71	0.81	0.83	0.71
		MSE	254207.28	254192.90	185063.61	165013.02	29005.58	79062.46	118460.06	67848.01

	RMSE	504.19	504.18	430.19	406.22	170.31	281.18	344.18	260.48
	nRMSE	14.11	14.11	12.04	11.37	4.67	7.58	10.86	8.39

4. CONCLUSION

Based on the findings of this study, it was infered that machine learning algorithms can accurately estimate cereal crop productivity. However, in this study, KNN and Ridge were shown to be the most accurate approaches for predicting rice crop yields, while ANN and KNN models for wheat crop were comparable. In the future, we can use meteorological data to estimate crop output in a big data environment. Thus, the study concludes that for the Shri Ganganagr district, KNN was the best model for district-level rice yield prediction while ANN was the best model for district-level wheat yield prediction.

REFERENCES

1. DeFries, R., Mondal, P., Singh, D., Agrawal, I., Fanzoand J., Remans, R. "Stephen Wood, Synergies and trade-offs for sustainable Agriculture. Nutritional yields and climate-resilience for cereal crops in Central India", *Global Food Security*.2016. 11: 44-53.
2. Reddy, S., Laha, G.S., Prasad, M.S., Krishnaveni, D., Castilla, N.P., Nelson, A., and Savary, S. Characterizing multiple linkages between individual diseases, crop health syndromes, germplasm deployment, and rice production situations in India, *Field Crops Research*. 2011.120(2): 241-253.
3. Venkateswarlu, B. Climate change, Adaptation and mitigation strategies in rain fed agriculture, *Journal of the Indian Society of Soil* Science. 2010.58: 27-35.
4. Deepa, L.R.A. and Praveen, N. Impact of Climate change and adaptation to green technology in India, Recent Advances in Space Technology Services and Climate Change.2010.RSTS & CC-2010), Chennai, India, 2010, pp.
5. Attri, S.D. and Rathore, L.S. Pre-harvest estimation of wheat yield for NW India using climate and weather forecast, *MAUSAM*. 2003. 54(3) 729–738,
6. Mishra, A., Rawat, S., Gautam, S., & Mishra, E. P. Comparison between Different Mustard Yield Prediction Models Developed using Various Techniques for Udaipur Region of Rajasthan. *International Journal of Environment and Climate Change*. 2022. 475–485. https://doi.org/10.9734/ijecc/2022/v12i1130997
7. Saseendran, A.S.K., Singh, K.K., Rathore, L.S., Singh, S.V. and Sinha, S.K. Effect of climate change in rice production in the tropical humid climate of kerala, India, *Climate Change.* 2000. 44: 495-514.
8. Reddy, V.R., and Pachepsky, Ya, A. Predicting crop yields under climate change conditions from monthly GCM weather projections. *Environmental Modelling & Software.* 2000.15(1): 79-86.
9. Pantazi, X., Moshou, D., Alexandridis, T., Whetton, R., & Mouazen, A. Wheat yield prediction using machine learning and advanced sensing techniques. *Computers and Electronics in Agriculture*. 2021. *121*, 57–65. https://doi.org/10.1016/j.compag.2015.11.018

10. Rehman, T. U., Mahmud, M. S., Chang, Y. K., Jin, J., & Shin, J. Current and future applications of statistical machine learning algorithms for agricultural machine vision systems. *Computers and Electronics in Agriculture*. 2019. *156*, 585–605. https://doi.org/10.1016/j.compag.2018.12.006
11. Elavarasan, D., Vincent, D. R., Sharma, V., Zomaya, A. Y., & Srinivasan, K. Forecasting yield by integrating agrarian factors and machine learning models: A survey. *Computers and Electronics in Agriculture*. 2018. *155*, 257–282. https://doi.org/10.1016/j.compag.2018.10.024
12. Soora, N.K., Aggarwal, P.K., Aggarwal, P.K., Saxena, R., Rani, S., Jain, S., & Chauhan, N.M. An assessment of regional vulnerability of rice to climate change in India. *Climatic Change.*2013.*118*, 683-699.
13. Das, B., Nair, B., V.K., and Venkatesh, P. Evaluation of multiple linear, neural network and penalized regression models for prediction of rice yield based on weather parameters for west coast of India. *International Journal of Biometeorology.* 2018. 62:1809-1822.
14. Johnson, M. D., Hsieh, W. W., Cannon, A. J., Davidson, A., & Bédard, F. Crop yield forecasting on the Canadian Prairies by remotely sensed vegetation indices and machine learning methods. *Agricultural and Forest Meteorology*. 2016.*218–219*, 74–84. https://doi.org/10.1016/j.agrformet.2015.11.003
15. Wu, Z., Huang, N. E., Long, S. R., & Peng, C. K. On the trend, detrending, and variability of nonlinear and non stationary time series. *Proceedings of the National Academy of Sciences*. 2007. *104*(38), 14889–14894. https://doi.org/10.1073/pnas.0701020104
16. Setiya, P. and Nain, A.S. Development of yield prediction model of rice crop for hilly and plain terrains of Uttarakhand. *Journal of Agrometeorology*. 2021.*23*(4), 452–456. https://doi.org/10.54386/jam.v23i4.162
17. Aravind, K.S.,Vashisth, A., Krishnan and B. Das. Wheat yield prediction based on weather parameters using multiple linear, neural network and penalized regression model. *Journal of Agrometeorology*. 2022. 24(1): 18-25.
18. Ghosh, R., Balasubramanian, S., Bandopadhyay, N., Chattopadhyay, K.K., Singh, K.K. and Rathore, L.S., 2014. *Development of Crop Yield Forecast Models under FASAL: A Case Study of Kharif Rice in West Bengal*. *Journal of Agrometeorology*, 16(1).
19. Verma, U., Piepho, H.P., Goyal, A., Ogutu, J.O. and Kalubarme, M.H. Role of climatic variables and crop condition term for mustard yield prediction in Haryana. *Int. J. Agric. Stat. Sci*, 2016. 12: 45-51.
20. Das, B., Nair, B., Arunachalam, V., Reddy, K.V., Venkatesh, P., Chakraborty, D. and Desai, S. Comparative evaluation of linear and nonlinear weather-based models for coconut yield prediction in the west coast of India. *International Journal of Biometeorology*. 2020. 64:1111-1123.
21. Vashisth, A., Singh, R., and Choudary, M. Crop Yield Forecast at Different Growth Stage of Wheat Crop using Statistical Model under Semi-Arid Region. *J. Agroecol.* 2014. 1(1):1-3.
22. Safa, M., Samarasinghe, S., and Nejat, M. Prediction of wheat production using artificial neural networks and investigating indirect factors affecting it: case study in Canterbury province, New Zealand, Journal of Agricultural Science and Technology. 2015.17(4):791-803.

23. Cattell, R. B. The screen test for the number of factors. Multivariate Behavioral Research. 1966. 1(2): 245-276.
24. Kaul, M., Hill, R. L., and Walthall, C. Artificial neural networks for corn and soybean yield prediction. *Agric. Syst*. 2005. 85: 1-18.
25. Kumar, S., Attri, S. D., and Singh, K. K. Comparison of Lasso and stepwise regression technique for wheat yield prediction. *J. Agro meteorology*. 2019. 21(2): 188-192.
26. Setiya, P.; Satpathi, A.; Nain, A.S., and Das, B. Comparison of Weather-Based Wheat Yield Forecasting Models for Different Districts of Uttarakhand Using Statistical and Machine Learning Techniques. *J. Agrometeorology*. 2022. 24: 255–261.
27. Abbas, F., Afzaal, H., Farooque, A. A., and Tang, S. Crop yield prediction through proximal sensing and machine learning algorithms. *Agron.* 2020.10(7): 1046.
28. Harrison, O. Machine Learning Basics with the K-Nearest Neighbors Algorithm.2018. Retrieved March 23, 2020, from https://towardsdatascience.com.
29. Rawat S, Singh R. K, Nain A. S. Analyzing Spatial Pattern of Weather Induced Yield Variability in Indian Mustard for Formation of Homogeneous Zones in North Western Himalaya and Indo-Gangetic Plains of India. Curr Agri Res 2018; 6(3)
30. Brownlee, J. *Machine Learning Mastery With Python* (v1.4). 2016. http://14.139.161.31/OddSem-0822-1122/Machine_Learning_Mastery_with_Python_2016.pdf
31. Donges, N. A Complete Guide to the Random Forest Algorithm.2019. Retrieved March 30, 2020, from https://builtin.com/data-science/random-forest-algorithm.html.
32. Jamieson, PD., Porter, JR., and Wilson, DR. A test of the computer simulation model ARCWHEAT1 on wheat crops grown in New Zealand. F Crop Res. 1991. 27:337–350.
33. Singh, M. and Nain, A.S. Statistical and machine learning approaches to study weather-disease-mustard yield relationship under varying environmental conditions.2020 PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2459209/v1]
34. Uno, Y., Prasher, S.O., Lacroix, R., Goel, P.K., Karimi, Y., Viau, A. and Patel, R.M. Artificial neural networks to predict corn yield from Compact Airborne Spectrographic Imager data. *Computers and electronics in agriculture*, 2005. *47*(2):149-161.