agronomy

MDPI

Article

Exploration of Machine Learning Approaches for Paddy Yield Prediction in Eastern Part of Tamilnadu

Vinson Joshua [1,*], Selwin Mich Priyadharson [1] and Raju Kannadasan [2]

[1] Department of Electronics and Communication Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai 600062, India; selwinmich@gmail.com
[2] Department of Electrical and Electronics Engineering, Sri Venkateswara College of Engineering, Sriperumbudur 602117, India; kannan.3333@yahoo.co.in or kannadasanr@svce.ac.in
* Correspondence: vinson.joshua@gmail.com

Abstract: Agriculture is the principal basis of livelihood that acts as a mainstay of any country. There are several changes faced by the farmers due to various factors such as water shortage, undefined price owing to demand–supply, weather uncertainties, and inaccurate crop prediction. The prediction of crop yield, notably paddy yield, is an intricate assignment owing to its dependency on several factors such as crop genotype, environmental factors, management practices, and their interactions. Researchers are used to predicting the paddy yield using statistical approaches, but they failed to attain higher accuracy due to several factors. Therefore, machine learning methods such as support vector regression (SVR), general regression neural networks (GRNNs), radial basis functional neural networks (RBFNNs), and back-propagation neural networks (BPNNs) are demonstrated to predict the paddy yield accurately for the Cauvery Delta Zone (CDZ), which lies in the eastern part of Tamil Nadu, South India. The performance of each developed model is examined using assessment metrics such as coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), coefficient of variance (CV), and normalized mean squared error (NMSE). The observed results show that the GRNN algorithm delivers superior evaluation metrics such as R^2, RMSE, MAE, MSE, MAPE, CV, and NSME values about 0.9863, 0.2295 and 0.1290, 0.0526, 1.3439, 0.0255, and 0.0136, respectively, which ensures accurate crop yield prediction compared with other methods. Finally, the performance of the GRNN model is compared with other available models from several studies in the literature, and it is found to be high while comparing the prediction accuracy using evaluation metrics.

Keywords: artificial neural network (ANN); crop yield prediction; machine learning algorithm; general regression neural networks (GRNNs)

Citation: Joshua, V.; Priyadharson, S.M.; Kannadasan, R. Exploration of Machine Learning Approaches for Paddy Yield Prediction in Eastern Part of Tamilnadu. *Agronomy* **2021**, *11*, 2068. https://doi.org/10.3390/agronomy11102068

Academic Editor: Jiftah Ben-Asher

Received: 23 August 2021
Accepted: 10 October 2021
Published: 15 October 2021

1. Introduction

1.1. Background

Due to the proliferation of the global population and living standards, demand for food grains is predicted to upsurge by 60%, notably in the middle of the 21st century [1]. The present change in climatic conditions threatens the crop yield that raises the risk to the farmers and associated dependence. Considering this urgent need, sustainable crop prediction is mandatory through a forecasting system that can precisely evaluate the crop conditions, crop kind, and its yield [2]. Crop yield methods are time-dependent and nonlinear by nature due to the amalgamation of an extensive array of interrelated factors influenced by non-arbitration and exterior features [3]. Conventionally, farmers made the crop yield prediction based on their previous practices and reliable historical evidence to make essential cultivation decisions. Notably, statistical methods adapt several regression approaches to associate historical crop yields to historical weather statistics that can be used to create yield predictions under changed weather settings [4] such as the availability of water resources, rainfall, temperature, drought, etc. Due to the swelling accessibility

and enhanced quality of the observed historical data, statistical methods have a great scale of accuracy [5,6]. In addition, remote sensing—specifically, satellite and airborne multi-spectral scanning, photography, and video—enables precision weed management over the generation of sensible and precise weed maps [7]. Furthermore, recently developing machine learning (ML) algorithms have greater ability of statistical methods to discover weather–yield relations [8].

Machine learning (ML) approaches are used for crop prediction using several mathematical and statistical methods, namely artificial neural networks, fuzzy information networks, decision tree, regression analysis, clustering, principal component analysis, Bayesian belief network, time series analysis, and Markov chain model. The application of these machine learning techniques in crop cultivation shows more tremendous advantages due to the availability of many data from several resources to obtain hidden knowledge [9]. Considering the need for machine learning techniques, a wide range of literature surveys is essential to derive a novel proposition to predict the crop yield's accuracy further.

1.2. Existing Methods—ML Algorithms for Yield Prediction

The forecasting agriculture process plays a crucial role in yield prediction using several advanced methodologies. There are dozens of research works that have been already carried out to attain high accuracy of crop yield prediction. Some of the notable pieces of literature are illustrated below (Table 1):

Table 1. Existing literature report.

Ref No	Year	Methodologies	Inferences
[10]	2016	Weighted histograms regression	− Proposed the design strategy for selecting soybean varieties to exploit maximum yield in the best season based on the knowledge attained from heterogeneous historical data. − The outcomes with the existing regression algorithm proved that the proposed algorithm offered an optimal selection of seed varieties.
[11]	2016	Regression Analysis (RA)	− Focused on analyzing the environmental constraints that impact the crop yield, namely area under cultivation, annual rainfall, and food price index. − RA analyzed the factors and groups them into explanatory and response variables that aids in attaining a decision.
[12]	2017	Gaussian process component and spatio-temporal structure	− Presented a scalable, accurate, and inexpensive technique to forecast crop yields using accessible remote sensing statistics (open source). − The proposed scheme improved the accuracy of the yield prediction pointedly along with a novel dimensionality reduction technique.
[8]	2017	Generalized regression neural network and radial basis function neural network	− The suggested method forecasted the yield of potato crops that were sown in flat and rough regions. Among the two methods, a generalized regression neural network was greater accuracy.
[13]	2017	Improved genetic algorithm-back propagation neural network prediction algorithm	− Proposed algorithm used to advance the yield–irrigation water model for forecasting the yield for various irrigation schemes under subsurface drip irrigation. − It offered more precise predictions of the yield with an average error of about only 0.71%.

Table 1. *Cont.*

Ref No	Year	Methodologies	Inferences
[8]	2018	Remote sensing and machine learning algorithms	− Discussed research growths accompanied within the last fifteen years on machine learning-based methods for accurate crop yield prediction and compared with remote sensing methods. − Concluded that the fast developments in sensing tools and machine learning techniques could deliver cost-effective and wide-ranging resolutions for improved crop and decision making.
[14]	2018	Multiple linear regression and radial basis function artificial networks	− Demonstrated the applications of the proposed algorithm to compute the probability of working days. − Performance criteria were considered, such as RMSE, MAPE, and R^2. − Radial basis function offered the highest R^2 compared with multiple linear regressions.
[15]	2019	Aggregated rainfall-based modular artificial neural networks and support vector regression	− Predicted the extent of monsoon rainfall using modular artificial neural networks. − Predicted the extent of chief Kharif crops yielded considering the rainfall data and area using support vector regression.
[16]	2019	Hybrid particle swarm optimization imperialist competitive algorithm, support vector regression	− Evaluated the performance of a proposed method to forecast apricot yield and identified significant factors affecting the yield. − The proposed scheme offered relatively high accuracy of prediction (RMSE of 1.737 and 2.329 for training and testing data, respectively).
[17]	2019	Support vector regression, K-nearest neighbor, random forest, and artificial neural network	− Used the agricultural dataset to contain 745 cases; 70% of statistics are randomly nominated to train the model and 30% are used for testing the model to evaluate the predictive capability. − Among the four algorithms, random forest offered the best accuracy in prediction.
[18]	2019	Deep neural network (DNN)	− Compared various artificial intelligence models to attain the most excellent crop yield prediction for the Midwestern United States (US). − Notably, the DNN model performed well, and its optimization process ensured the most acceptable configurations for the drop-out ratio, layer structure, cost function, and activation function.
[19]	2019	Deep neural network (DNN)	− With the suggested scheme, greater prediction accuracy with a root mean square error (RMSE) of 12% of the average yield and 50% of the standard deviation for the validation dataset using predicted weather data.
[20]	2019	Artificial neural network	− Evaluated five different ANN methods, namely generalized feed-forward, multilayer perceptron, Jordan/Elman, principal component analysis, and radial basis function. − Among these models, multilayer perceptron offered the best prediction.
[21]	2019	Machine learning and big data	− Various machine learning algorithms were examined to verify the usefulness in predicting crop yield. − Prediction of crop yield using machine learning methods in big data computing pattern was demonstrated.

Table 1. *Cont.*

Ref No	Year	Methodologies	Inferences
[22]	2019	Support vector machine, random forest, and neural network	− Used the enhanced vegetation index from MODIS and solar-induced chlorophyll fluorescence from GOME-2 and SCIAMACHY as metrics to predict crop production. − The machine learning method offered the best yield prediction compared with the regression method.
[23]	2020	Hybrid genetic algorithm-based back-propagation neural network (GA-BPNN) model	− The proposed scheme was used to offer complimentary data on maize growth at the vital growth phase. − The hybrid concept enhances the yield significantly compared with the pure back-propagation scheme.
[24]	2020	Proximal Sensing (PS) and machine learning algorithms	− PS surveyed the soil and crop variables potentially for variations in yield. − Four algorithms were demonstrated: linear regression, elastic net, k-nearest neighbor, and support vector regression to forecast potato yield from soil and crop data properties collected over proximal sensing.
[25]	2021	Partial least squares and radial basis function neural network.	− Carried out to estimate the feasibility of using Vis/near-infrared spectroscopy to determine the potassium concentration and petioles of distinct variety and mixed lettuce leaves of two varieties. − Partial least squares offered R^2 of 0.83, residual predictive deviations of 1.95, and RMSE of 39.07.

1.3. Objectives

Considering the above inferences, crop yield prediction needs a more accurate and reliable method to attain more precision using evaluation metrics. Based on these needs, this work focused on the following objectives:

- To assess the paddy crop yield data from high potential real-time locations.
- To estimate the crop yield prediction using a statistical model (MLR).
- To demonstrate advanced machine learning techniques such BPNNs, RBFNNs, GRNNs, and SVR for crop yield prediction.
- To analyze the adapted machine learning techniques using evaluation metrics such as R^2, RMSE, MAE, MSE, MAPE, CV, and NSME.
- To select and recommend the best accurate prediction technique to evaluate the crop yield.

2. Data Collection

The historical data (paddy crop) of the CDZ, which lies in the eastern part of Tamil Nadu, South India, is considered for this study. The CDZ has a total geographic land area of 14.47 lakh hectares. It covers several districts of Tamilnadu namely Thanjavur, Thiruvarur, Nagapattinam, Trichy, Ariyalur, Cuddalore, and Pudukkottai districts (Figure 1). In this zone, paddy is the principal crop. In the rice-based cropping system, it is either single or double cropped. In this work, 50 fields in the Thirichirapalli, Perambalur, and Pudukottai districts in CDZ (Figure 1) are collected for two seasons (June 2018–September 2018 and October 2018–January 2019.). There are two main reasons for selecting these three regions: the foremost percentage of paddy yield is harvested in these regions of Tamilnadu; the soil types plays critical role i.e., Thirichirapalli mostly has alluvial soils, Perambalur has mostly heavy clay soils, and Pudukottai consists of alluvial and laterite soils.

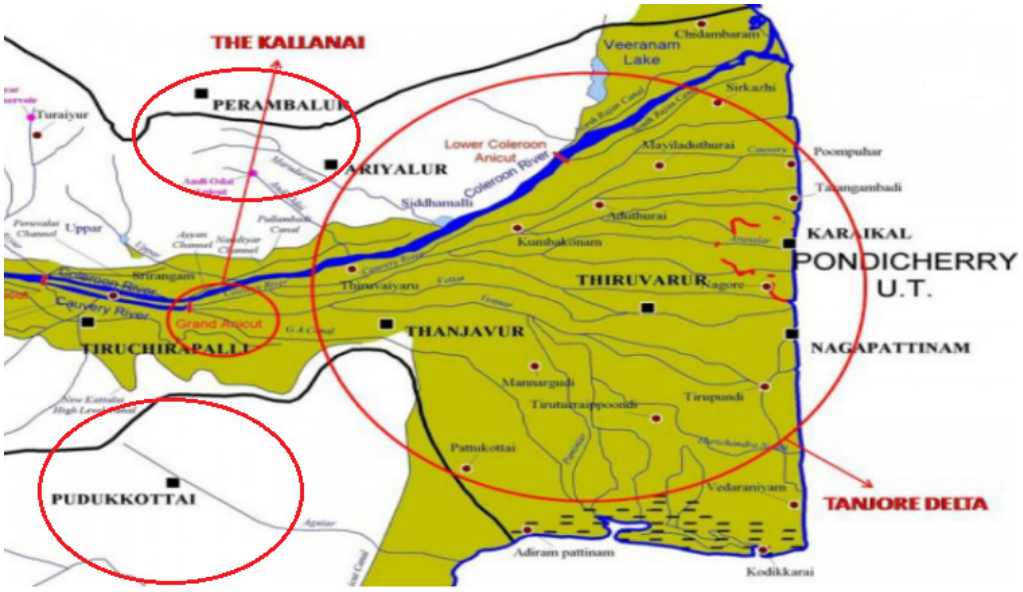


Figure 1. CDZ belt in eastern part of Tamilnadu.

To get a better overview of the independent variables (features), they can be grouped into soil information (pH value), humidity (rainfall), solar information (temperature), nutrients (nitrogen, phosphorus, and potassium), and field management (Urea). The data were collected from the meteorological department of India [26], agricultural department of Tamilnadu [27], and the statistical department of Tamilnadu [28]. The complete description of the considered parameters is illustrated in Table 2.

Table 2. Description of the selected sites.

Parameters	Tiruchirappalli	Pudukkottai	Perambalur
pH range	8.2–9.6	6.8–8.5	8.09–8.6
Temperature	24–38	24–33	25–34
Mean annual rainfall	761	821	861
SW monsoon (June–September): mm	273.3	351.9	270
NE monsoon (October–December): mm	394.8	394.1	466
Field	16	21	13

Furthermore, the data collection of the selected sites such as mean rainfall, temperature, fertilizer, nitrogen, phosphorous, potassium, pH value, and yield are obtained from the digital sources. A total of 280 samples are analyzed initially, and the repeated and insignificance data are merged to attain 100 rows of data. From the finalized data, a minimum, maximum, mean, and standard deviation are attained and illustrated in Table 3.

The central part of data collection is creating a training network that can forecast the atmospheric components, namely temperatures, rainfall, etc., for a specific station. Rainfall is a significant factor of agriculture production, and its dissimilarity can affect crop production. The temperature is one of the essential factors of the metrological parameter that supports any crop growth. In this work, data are collected from online sources such as data.gov.in and indiastat.org. The datasheets are prepared based on the retrieved sources for analysis. Notably, this work adapted annual abstracts about a crop for two periods in a year. The input datasets are prepared with several samples and arranged in an Excel sheet and later loaded into MATLAB for analysis. Among the loaded datasheet, 70% of

the datasets are used for training, and 30% of the dataset is considered for testing. The test data offer an independent degree of neural network performance employing MSE.

Table 3. Data collection for yield prediction from selected sites.

Variables	Rows	Minimum	Maximum	Mean	Std. Deviation
Mean Rainfall (mm)	100	266.0	464.0	366.4	75.59
Temperature (°C)		24.0	38.0	31.5	4.40
Fertilizer(urea) (kg/ha)		123.50	197.6	166.62	24.86
Nitrogen (N) (kg/ha)		143.26	197.6	174.13	16.66
Phosphorus (P)(kg/ha)		44.46	61.75	52.04	4.75
Potassium (K) (kg/ha)		37.05	54.34	44.48	4.49
pH value		6.90	8.93	8.12	0.48
Yeild (kg/ha)		2358.0	3189.0	2773.5	207.7

To attain the designated objectives, the following steps need to demonstrate the application of the selected machine learning algorithms.

Step 1: Collect the data using available sources.
Step 2: Distribute the data into two segments: training data (70%) and testing data (30%).
Step 3: Develop the machine learning model to assess the crop yield.
Step 4: Predict the crop yield using adapted techniques.
Step 5: Determine the evaluation metrics for each model.
Step 6: Recommend the best-rated technique for crop yield using observed outcomes.

3. Methodology

3.1. Statistical Analysis

Statistical analysis is adapted primarily, namely multiple linear regression (MLR), to determine the effect of some independent variables on dependent variables to compute the linear dependence of the variables [29]. It defines an association between known (x) and unknown variables (y) based on the random noise and its parameters, and it is expressed as below:

$$y_i = \beta X_i + \varepsilon_i \quad (1)$$

where y_i denotes a predicted rate; $X_i = (1, x_1, x_2, x_3, \ldots x_n)$ are the terms for the explanatory vector variables; $\beta = (\beta_0, \beta_1, \beta_3, \ldots \beta_k)^T$ represents a vector coefficient; ε_i denotes a random error for ith observation.

3.2. Machine Learning Techniques

Soft computing is a collection of practices applied in many fields and falls under several computational intelligence categories. It includes fuzzy systems (FS), evolutionary computation (EC), artificial neural network (ANN), probabilistic reasoning (PR), etc. As stated earlier, the ANN model is adapted in this work to determine the performance for crop yield prediction, namely support vector machine (SVM), generalized regression neural network (GRNN), radial basis function neural network (RBFNN), and back-propagation neural network (BPNN). There are seven input parameters: rainfall, fertilizer, temperature, nitrogen, phosphorus, potassium, soil pH, and one output can be obtained, likely crop yield. The complete process is carried out in MATLAB 2018(b) software to implement the models based on the proposed algorithm.

Further, normalization is adapted to prepare data reduction and remove the data redundancy for machine learning applications. It aids in amending the numeric columns in the specified dataset into a standard scale without deforming in their ranges. Generally, it must lie in the data range of 0 and 1, which is essential before applying to any soft computing models. Typically, three types of normalization techniques are used: Min–Max

normalization, Z-score normalization, and decimal scaling. In this work, the Min–Max normalization technique is considered for data preparation using the following equation:

$$\text{Normalized } (D) = N = \frac{D - \text{Min}(P)}{Max(P) - Min(P)} \quad (2)$$

where Min (P) and Max (P) indicate the minimum and maximum value of attribute P, respectively.

3.2.1. Support Vector Machine (SVM)

SVM is a novel supervised computational machine learning method for classification and regression that depends on statistical learning theory advancements (Figure 2), and the required input parameters are shown in Table 4. It can train nonlinear models based on the principles of structural risk minimization (SRM) that minimize an upper bound on the generalization errors rather than empirical error minimization as implemented in neural networks [15]. It was realized based on the Vapnik–Chervonenkis theory (VC) conventionally and emerged as a general mathematical framework recently to determine dependencies from finite sample sets. This theory integrates fundamental concepts with associated learning principles, precise formulation, and a self-consistent mathematical model.

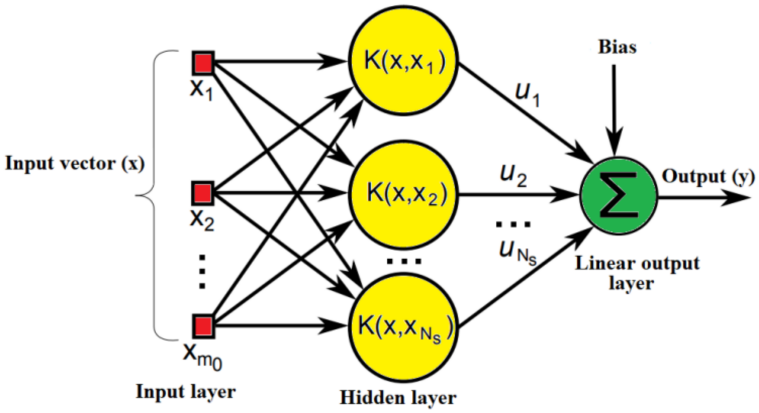


Figure 2. Architecture of SVM.

Table 4. Input parameters of SVM.

Parameters	Descriptions/Values
Type of SVM model	Epsilon-SVR
SVM kernel function	Radial basis function (RBF)
Search criterion	Minimize total error
Number of points evaluated during search	1093
Minimum error found by search	0.462196
Epsilon	0.001
C	34.5930771
Gamma	0.41179479
P	0.21545292
Number of support vectors	73

3.2.2. Generalized Regression Neural Network (GRNN)

Donald F. Specht proposed GRNN with a variation of the radial basis function neural network (RBF) in 1991. It is a one-pass neural network with highly parallel construction [30].

It is an algorithm based on function approximation (estimation) and a statistical technique named kernel regression. GRNN can be trained very quickly, and data propagated forwarded only once, unlike other neural network algorithms. The desired output can be determined by considering an average of assigned weights of the training output data set. The weight of each result can be calculated using the Euclidean distance function between the training and testing data. If the Euclidean distance is more than the total weight, the output is less than the additional weight and they should be assigned to the output.

GRNN comprises four layers: an input layer, a pattern layer, a summation layer, and an output layer (Figure 3). The size of input neurons in the input layer depends on the total number of the experimental parameters. The input layer feeds the input to the pattern layer, and each neuron presents a training pattern and output. The primary purpose of the pattern layer is to calculate the Euclidean distance along with the activation function and forward it to the summation layer. The summation layer has two sub-parts: a numerator (N) and denominator (D). The numerator part consists of the addition (summation) of the multiplication of training output data and activation function, and the denominator part has the acquisition of all specified activation functions. This summation layer feeds both the numerator and denominator parts to the output layer.

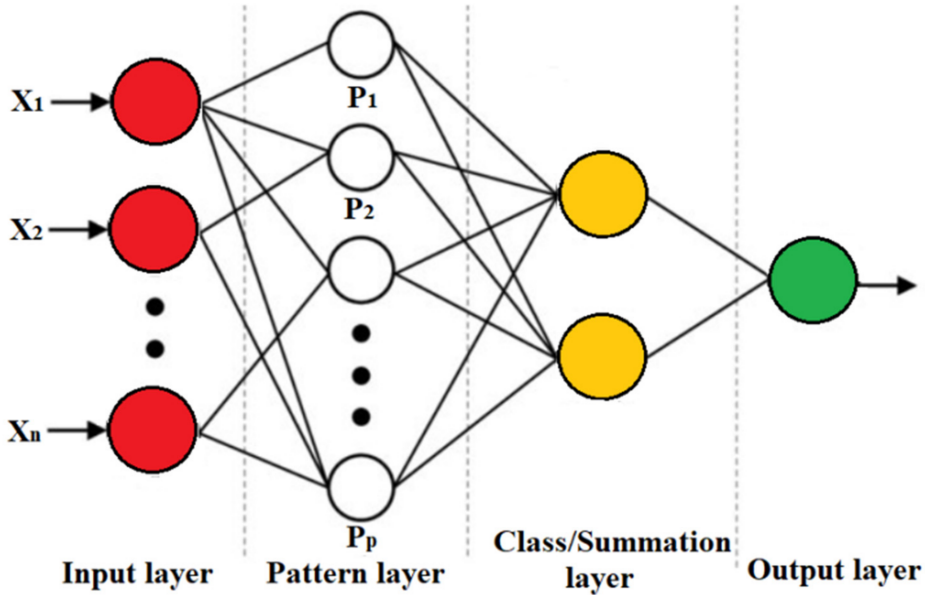


Figure 3. GRNN architecture.

3.2.3. Radial Basis Functional Neural Network (RBFNN)

An ANN adapts RBF as an activation function such as the input layer, hidden layer, and linear output layer. It is derived from the concept of function approximation, which is a well-known and popular alternative model to the MLP that has a more straightforward structure and a quicker training process [14]. Basically, it is used to detect the minimum number of hidden layers or perceptions in a single hidden layer until a minimum error value is stretched. The input layer has nodes that match the number of datasheet input parameters. An invisible layer found its response using a radial basis function in every perceptron. In general, a Gaussian function and an output layer create a linear weighted sum of hidden neuron outputs and supply the response to the network. The structure of the RBFNN network is depicted in Figure 4 and the required input parameters are illustrated in Table 5.

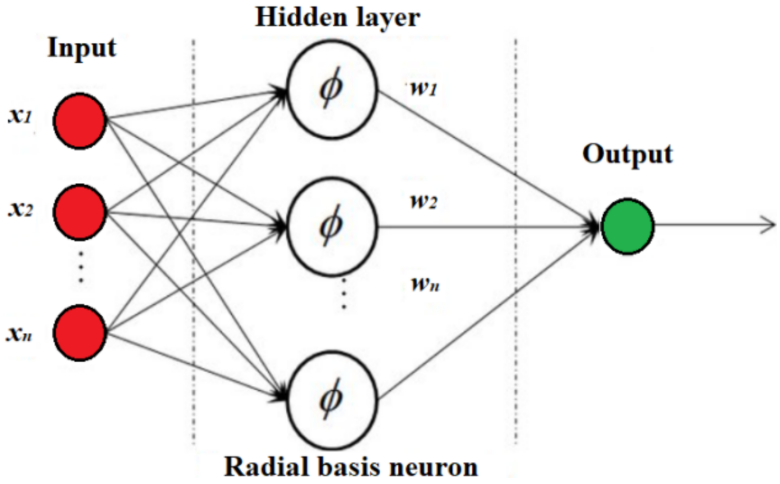


Figure 4. Network architecture of RBFNN.

Table 5. Input parameters of RBFNN.

Parameters	Ranges/Values
No. of neurons	25
Minimum radius	0.019
Maximum radius	395.265
Minimum lambda	0.06458
Maximum lambda	8.64019
Regularization lambda (final weights)	1.549×10^{-5}

3.2.4. Back Propagation Neural Network (BPNN)

An ANN adapts several layers that can approximate multifaceted mathematical functions to process the data. BPNN is the most broadly adapted algorithm for an ANN application that takes an error gradient as a back-propagation [13]. In this work, the proposed BPNN algorithm is considered to adjust the simulated value to attain more crop prediction accuracy. It comprises four different stages: initialization of weights, feed-forward, back-propagation of errors, and updating of weights and biases. The comprehensive architecture of the proposed model is depicted in Figure 5 and the input, hidden, and output parameters are given in Table 6.

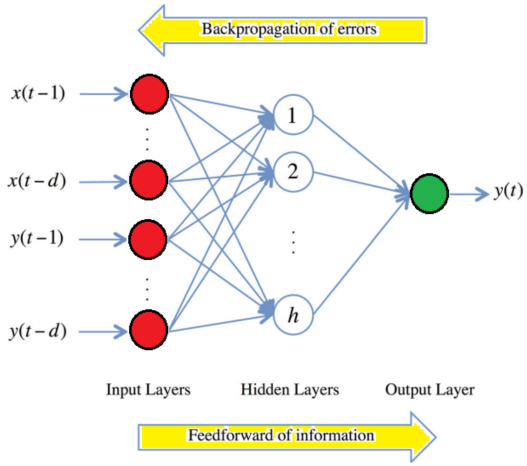


Figure 5. Network architecture of BPNN.

Table 6. Input parameter of BPNN.

Layer	Neurons	Activation
Input	7	Pass through
Hidden	15	Logistic
Output	1	Linear

4. Model Performance

Different standard statistical performance evaluations evaluate various conventional predictor model performances. The most widely used statistical measures are coefficient of determination (R^2), root mean square error (RMSE), mean absolute error (MAE), mean squared error (MSE), mean absolute percentage error (MAPE), coefficient of variance (CV), and normalized mean squared error (NMSE). The derivative functions of such parameters are given in the following equations [5,31–33]:

$$R^2 = \left\{ \frac{1}{N} * \frac{\sum(X_i - \overline{X}) * (Y_i - \overline{Y})}{(\sigma_X - \sigma_Y)^2} \right\}^2 \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=0}^{n} |A_i - P_i|^2}{n}} \tag{4}$$

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} \left| Y_i - \overrightarrow{Y}_i \right| \tag{5}$$

$$\text{MSE} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \overrightarrow{Y}_i \right)^2 \tag{6}$$

$$\text{MAPE} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{Y_i - \overrightarrow{Y}_i}{Y_i} \right| \times 100 \tag{7}$$

$$\text{CV} = \frac{S_i}{Y_i} \tag{8}$$

$$\text{NMSE} = \frac{\| (Y_i - \overrightarrow{Y}_i) \|_2^2}{\| \overrightarrow{Y} \|_2^2} \tag{9}$$

where A_i and P_i are measured and predicted values, respectively; N is the number of observations; X_i and Y_i are the X and Y value of observation 'i', respectively; $\overline{X}$ and $\overline{Y}$ are the mean X and Y, respectively; σ_x and σ_y are the standard deviations of X and Y, respectively; and S_i represents an intertemporal variance.

5. Results and Discussions

In this section, statistical and proposed machine learning models are demonstrated in a virtual platform. The statistical analysis adapts several components: rainfall, fertilizer, temperature, nitrogen, phosphorous, and potassium. The higher correlation and lower error scale model will be considered the best technique for crop yield (kg/acre) prediction.

5.1. Statistical Analysis

The first case represents the outcome of the statistical approach, namely MLR that offers a multiple R and R^2 of about 0.9427 and 0.8888 (Table 7). Then, the adjusted R^2 and standard deviation are observed as 0.8803 and 0.6862, respectively (Table 3). The crop yield (Q/ha) between prediction and measured data is depicted in Figure 6. Moreover, MSE and RMSE metrics offer an average range likely of 0.5247 and 0.6586, respectively.

However, these outcomes are not excellent, because MLR is the method that uses simple linear association among a dependent and independent variable. This technique adapted a least squares model, which is simple in design, but outcomes are not great. This method offers moderate results for developing models to restructure climate variables from tree ring services (error percentage is higher i.e., −14% and +13%) [34] but not shown potential fallouts for crop yield prediction. Therefore, the crop yield prediction can be further improved using machine learning techniques as illustrated in subsequent sections.

Table 7. Implementation and outcomes of MLR method.

Regression Statistics								
Multiple R	0.942762							
R Square	0.8888							
Adjusted R Square	0.88034							
Standard Error	0.682364							
Observations	100							
ANOVA								
	df	**SS**	**MS**	**F**	**Significance F**			
Regression	7	8.1039	1.15770	105.0487	4.69E–41			
Residual	92	1.0138	0.01102					
Total	99	9.1178						
	Coefficients	**Standard Error**	**t Stat**	***p*-Value**	**Lower 95%**	**Upper 95%**	**Lower 95.0%**	**Upper 95.0%**
Intercept	0.439148	0.11923	3.683201	0.000389	0.202347	0.675948	0.202347	0.675948
Mean Rainfall (mm)	0.04361	0.109511	0.398225	0.691387	−0.17389	0.261109	−0.17389	0.261109
Temperature (°C)	−0.36972	0.106524	−3.47074	0.000792	−0.58128	−0.15815	−0.58128	−0.15815
Fertilizer(urea) (kg/ha)	−0.13005	0.092188	−1.41074	0.161694	−0.31314	0.05304	−0.31314	0.05304
Nitrogen (N) (kg/ha)	0.343809	0.094175	3.650756	0.000434	0.15677	0.530848	0.15677	0.530848
Phosphorus (P) (kg/ha)	0.112423	0.072317	1.554591	0.123477	−0.0312	0.256051	−0.0312	0.256051
Potassium (K) (kg/ha)	0.304443	0.079279	3.840153	0.000226	0.146988	0.461897	0.146988	0.461897
pH value	−0.04314	0.049602	−0.86974	0.386708	−0.14165	0.055373	−0.14165	0.055373

5.2. Machine Learning Techniques

Several studies demonstrated the prediction of paddy yield using machine learning methods [35–39]. However, there is a need to enhance the prediction accuracy for reliable crop yield. As discussed above, R^2, RMSE, MAE, MSE, MAPE, CV, and NMSE metrics are applied to evaluate the accuracy of the proposed ANN algorithm, such as SVM, GRNN, RBFNN, and BPNN for crop yield prediction. Furthermore, each ANN model generates a plot that represents the crop yield prediction against original yield data (Figure 7). From the illustrations, all the metrics are assessed using the above-mentioned formulas. Then, the metrics are evaluated to test the accuracy of the considered algorithms between predicted and original crop yield.

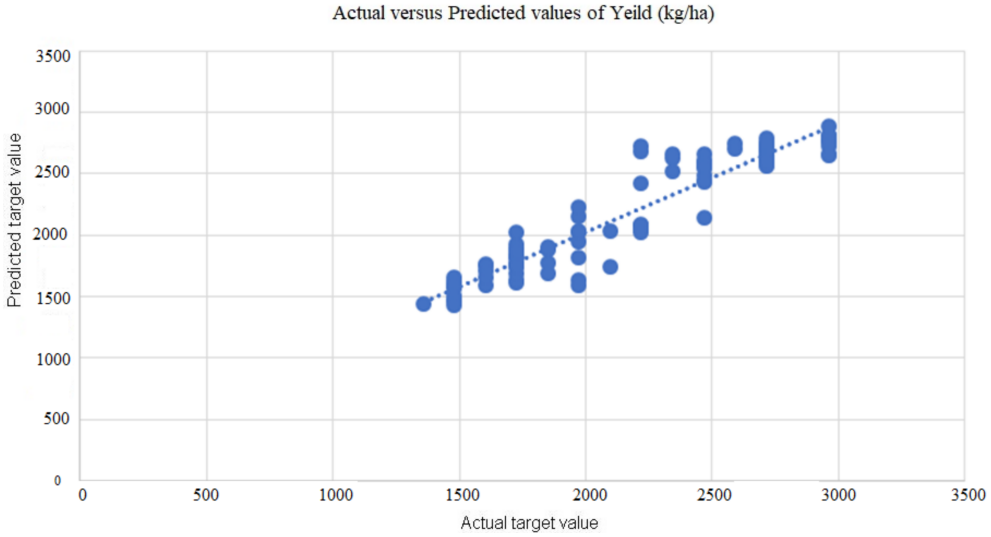
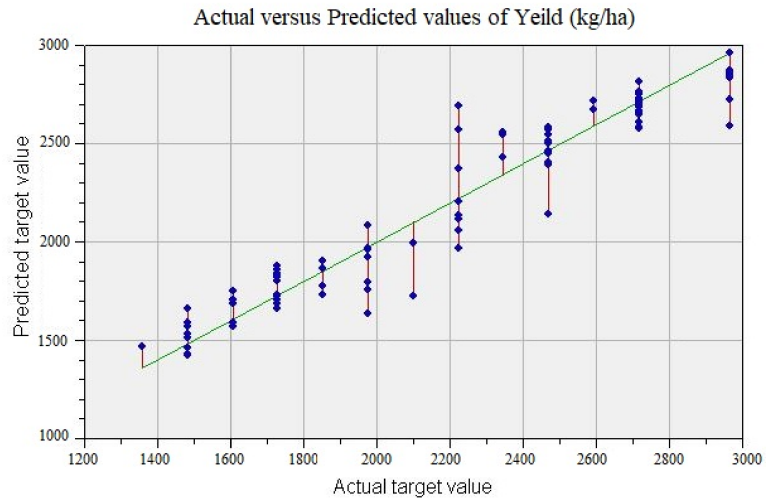
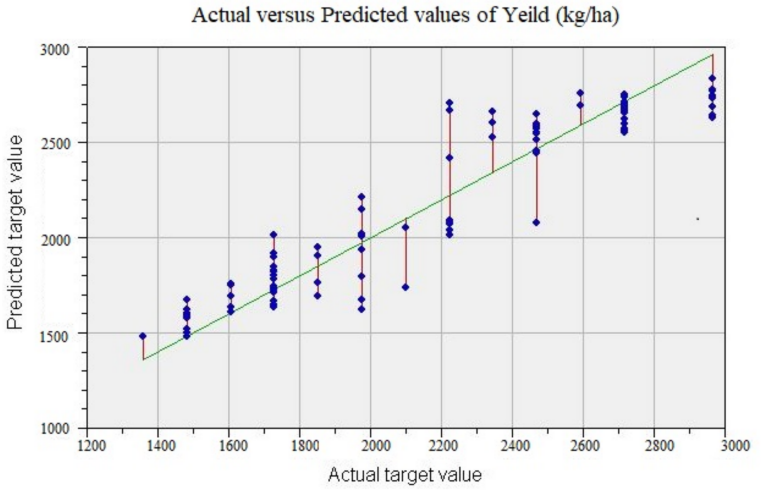


Figure 6. MLR model.

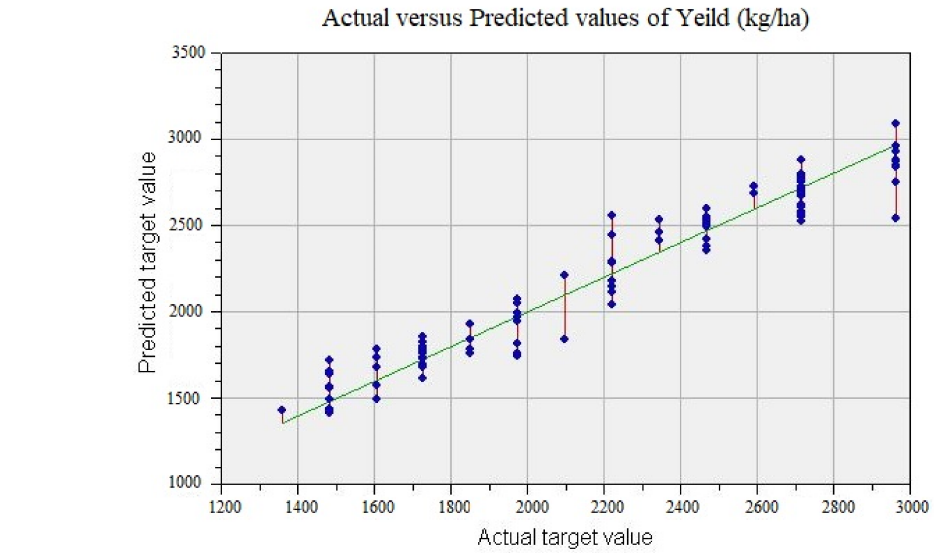


(**a**)

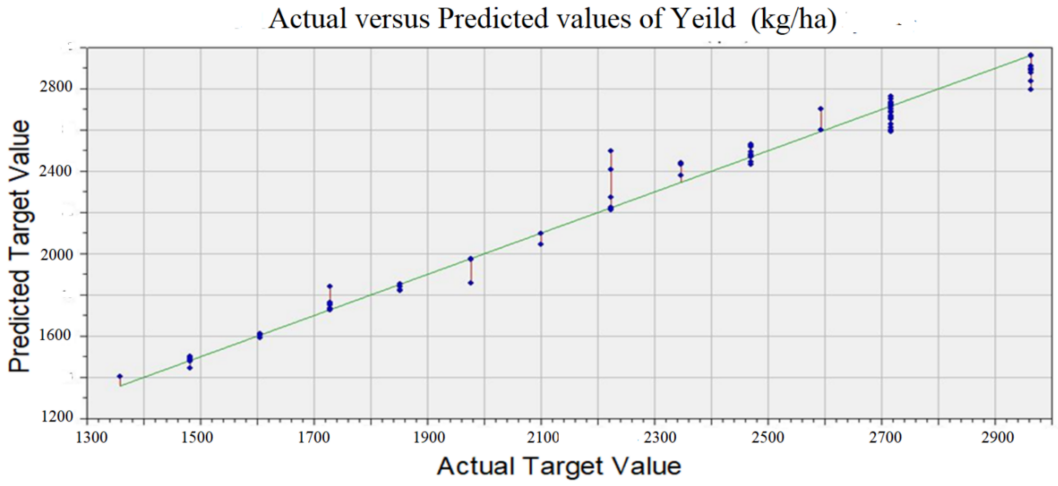


(**b**)

Figure 7. *Cont*.



(**c**)



(**d**)

Figure 7. Correlation of determination (R^2): (**a**) SVM, (**b**) BPNN, (**c**) RBFNN, (**d**) GRNN.

The effectiveness of the proposed results is compared with the recent pieces of literature. Notably, Elavarasan et al. suggested deep reinforcement learning to develop the prediction scheme [35]. It was noted that the efficiency and accuracy of the proposed scheme were sophisticated compared with other models, likely LSTN (Long Short-Term Network), BAN (Big Ass Number), and RAE (Regularized Auto Encoder). However, limited evaluation metrics were considered for precision prediction for paddy cultivation. Notably, CV and NMSE were not adapted to ensure the better precision of the proposed model.

Furthermore, Gopal et al. [36] designed a hybrid model such as MLR-ANN for crop yield prediction. In this work, MLR's coefficients and their bias were engaged in initializing. The suggested hybrid model displayed improved prediction precision compared with SVR, K-NN (K-nearest neighbors), and RF (random forest). The precision evaluation metrics of the SVM showed a better result compared with BPNN and RF. However, there was little consideration of the evaluation metrics that require detailed evaluation to ensure the effectiveness of the suggested algorithms. Some of the work proposed the novel algorithms such as Hybrid CNN-RN, MARS, and DNN for corn and soybean yield prediction. However, only RMSE and correlation coefficient evaluation metrics are considered for prediction [18,40]. For paddy yield prediction, few works proposed RF, MLR-ANN, and DT. However, the evaluation metrics MAE, RMSE, and R were considered [22]. Furthermore,

the researcher carried out tomato yield prediction [41], but only the RMSE metric was considered. In addition, a prediction of palm yield was proposed using genetic algorithm, but only R^2 and MSE were considered for evaluation [42]. Additionally, wheat and barley yield predictions were proposed using the CNN algorithm; however, only the MAPE metric was considered [43]. Consolidating all these inferences, the adaptation numbers of evaluation metrics are not great, and therefore, this work focused on computing the wide range of evaluation metrics such as R^2, RMSE, MAE, MSE, MAPE, CV, and NMSE using the proposed algorithms to ensure their effectiveness. In addition, paddy yield prediction using SVM, RBFNN, GRNN, and BPNN are not demonstrated by the researchers remarkably.

To simplify the comparative analysis between various algorithms, individual metrics are presented in Figure 8, namely R^2, RMSE, MAE, MSE, MAPE, CV, and NMSE. As stated above, higher accuracy represents greater R^2 (closer to unity) and lower RMSE, MAE, MSE, MAPE, CV, and NMSE. In line with this statement, it is proved that the GRNN algorithm performed well compared with other ANN and statistical methods. Notably, the R^2 metric of GRNN attained a more excellent value of about 0.9863, which is far better than other methods (Figure 8a).

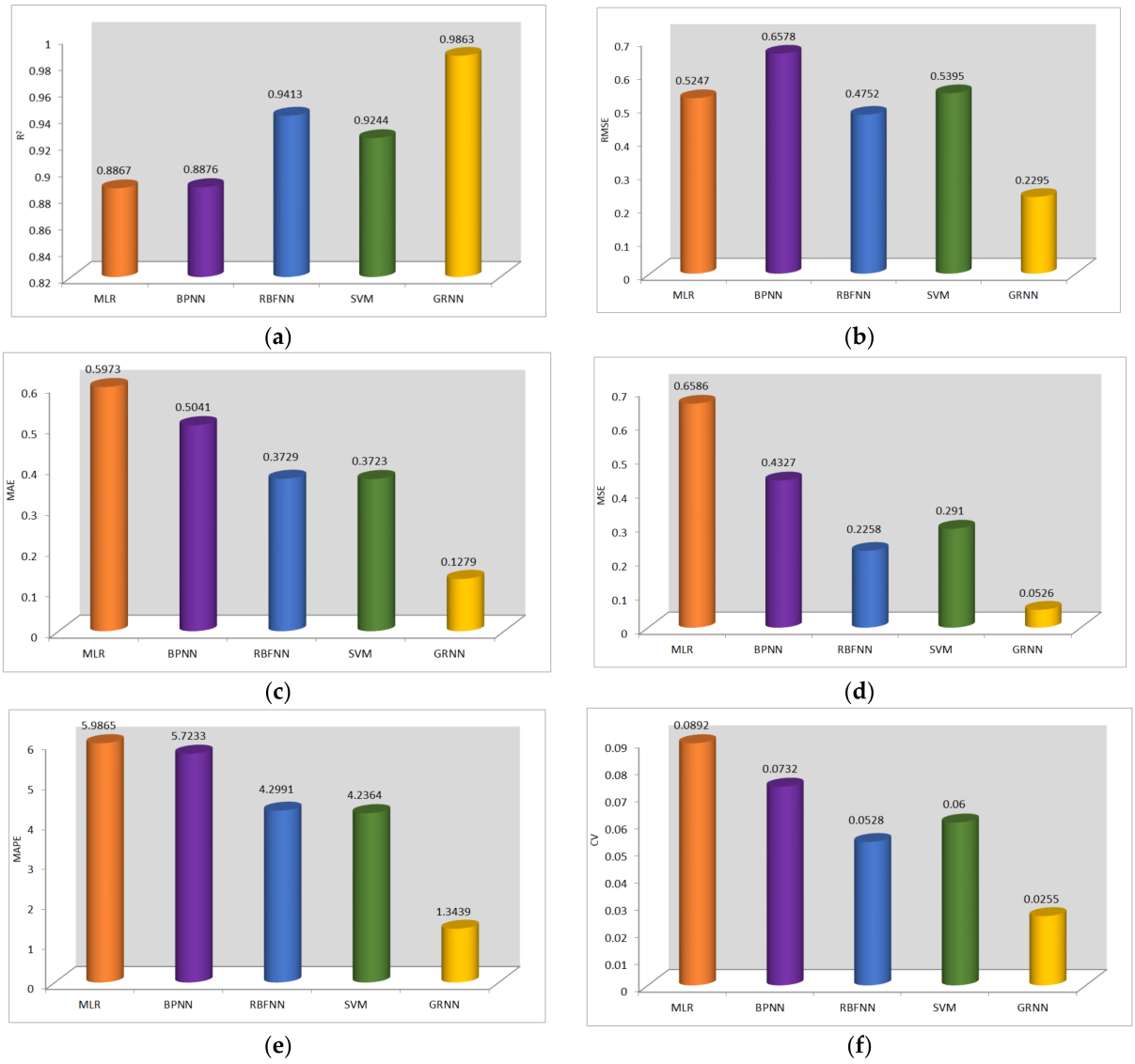
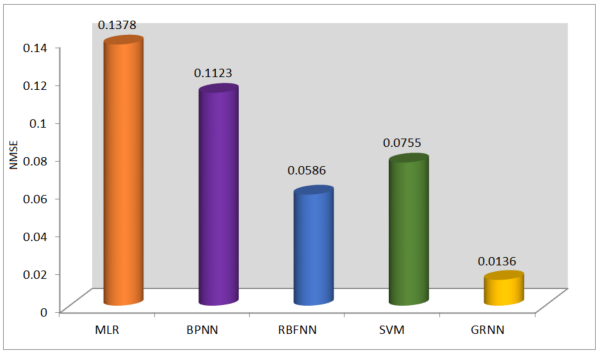


Figure 8. *Cont.*



(**g**)

Figure 8. Comparative analysis between different techniques: (**a**) R^2, (**b**) RMSE, (**c**) MAE, (**d**) MSE, (**e**) MAPE, (**f**) CV, (**g**) NMSE.

Furthermore, it is perceived that the RMSE metric of GRNN shows a lower value of about 0.2295 (Figure 8b), representing the accuracy of the crop yield compared with other adapted schemes. Similarly, the metrics of MSE, MAE, MAPE, CV, and NSME show the least range for GRNN models: about 0.1279, 0.0526, 1.3439, 0.0255, and 0.0136, respectively which are superior compared with other methods such as MLR, SVM, RBFNN, and BPNN (Figure 8b–f). All these outcomes attest the accuracy of the paddy yield prediction on the CDZ zones, i.e., the eastern part of Tamilnadu.

As the considered metrics show the higher effectiveness of the GRNN algorithm, it is essential to compute the running time of all the adopted ANN models. Therefore, the individual run times of the models are computed and illustrated in Figure 9. It is observed that the GRNN model completed the prediction task within 880 ms, which is comparatively lower than other ANN models such as SVM, BPNN, and RBFNN. This high-speed computation is not possible with other numerical models due to the complicated mathematical models that tend to increase the inaccuracy of the prediction. In addition, the numerical model has greater limitations regarding the number of input parameters, which is not a concern for the ANN model.

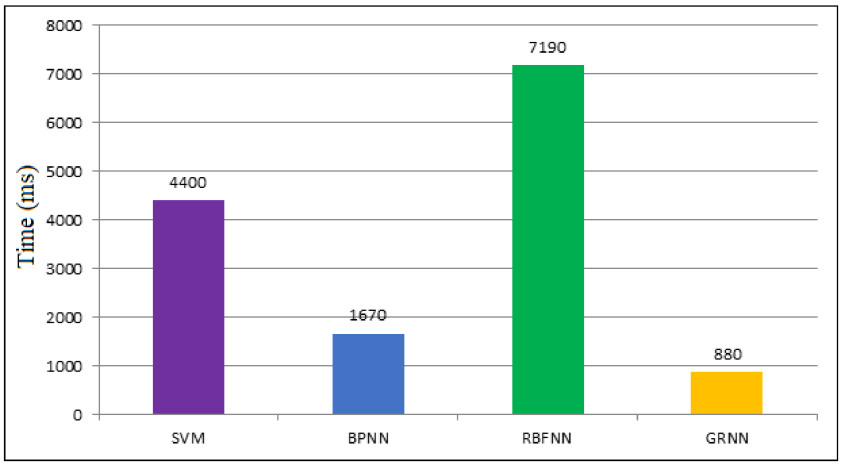


Figure 9. Run time of the ANN models.

Consolidating all the inferences and statements, it is perceived that the ANN algorithms have done well for crop yield prediction. Notably, the GRNN algorithm offered superior results compared with other adapted techniques such as SVM, BPNN, and RBFNN using three performance metrics. Furthermore, the prediction accuracy of the GRNN model is compared with other competitive methods from the literature. Notably, the regression analysis model was adopted by the authors for crop yield prediction accuracy, and the coefficient R^2 attained a maximum scale of about 0.7272 [30]. In addition, the same co-

efficient was evaluated using the particle swarm optimization–imperialist competitive algorithm–support vector regression (PSO-ICA-SVR) method, and it attained the best value of 0.874 [17]. In addition, the performance of the random forest method was considered for crop yield prediction using the R^2 coefficient, and it obtained better results, i.e., 0.92 [6]. Comparing these inferences, the accuracy of the proposed GRNN model shows extremely good scale of about 0.9863 (about 7.53% is increased compared with the random forest method). Furthermore, other evaluation metrics such as RMSE and MAE are compared with existing methodologies; the PSO-ICA-SVR model offered minimum RMSE and MAE of about 1.418 and 1.737 respectively 17]. However, the proposed GRNN model shows the lowest values of about 0.2295 (RMSE) and 0.1279 (MAE). Other evaluation metrics (MSE, MAPE, CV, and NMSE) are not demonstrated greatly by the researcher using machine learning models. This work targeted all possible evaluation metrics to validate the effectiveness of the proposed model.

Moreover, the absolute yield of the selected location is compared with other parts of Indian states, and the complete comparative case is illustrated in Figure 10. It is found that the state of Tamilnadu attained the highest yield: about 3191 kg/ha [44]. This is owing to the optimum parameters of the state: notably, a mean temperature of 28 °C, higher rainfall of 464 mm (3 months), and pH value about 6.9. In addition, paddy cultivation parts of Tamilnadu comprise a wide range of alluvial soil, which is suitable for paddy cultivation. The predicted values of the machine learning model almost match with the absolute yield of the Tamilnadu but with different accuracy based on the effectiveness of the individual algorithm. It is already stated that the accuracy of the GRNN model shows better scale among other selected machine learning models.

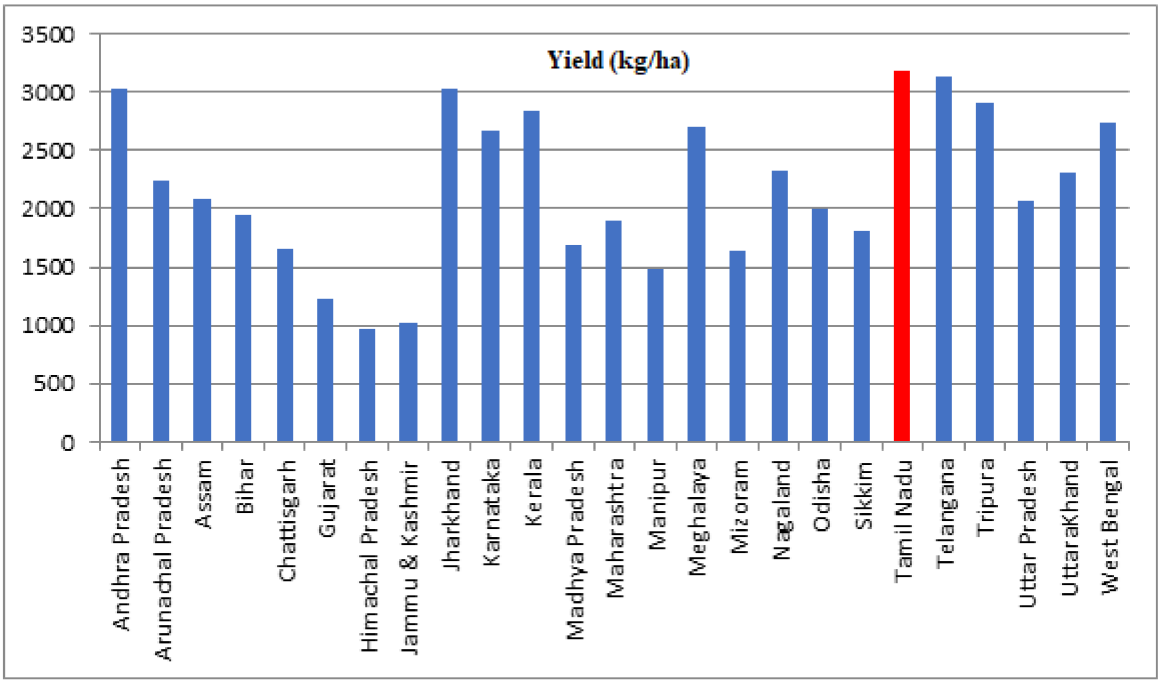


Figure 10. Comparative study of yield among Indian states.

These research findings confirm the consistency between predicted yields and the government's yield statistics. As per the literature survey, there are no benchmark data sets available for crop yield, and it is challenging to predict owing to diverse biological parameters. Therefore, GRNN can be adapted for the crop yield prediction for effective outcomes that can reduce the risk factor for the farmers.

6. Conclusions

Prediction of crop yield is carried out using statistical and machine learning algorithms. Specifically, the statistical study of likely MLR techniques and machine learning algorithms such as SVM, GRNN, RBFNN, and BPNN are considered for evaluation to attain crop yield prediction of higher accuracy. Model performance metrics are adapted to scrutinize the accuracy level of the different algorithms. With the observed outcomes, the following conclusions are made:

- Machine learning algorithms attained exceptionally greater yield prediction accuracy than statistical methodology based on the results of evaluation metrics.
- Among the four machine learning algorithms such as SVM, RBFNN, GRNN, and BPNN, GRNN predicted the yield more precisely.
- R^2, RMSE, MAE, MSE, MAPE, CV, and NSME performance metrics of GRNN showed a better scale of 0.9863, 0.2295, 0.1290, 0.0526, 1.3439, 0.0255, and 0.0136, respectively.
- Run time of the GRNN model shows a superior scale of 880 ms, which is comparatively less than that of the other ANN models.
- Compared with other existing models from the literature reports, the R^2 metrics of the proposed model (GRNN) are improved by 7.53%.
- The absolute yield of Tamilnadu and other Indian states are compared, and it is found that Tamilnadu acquired the highest yield, about 3191 kg/ha, and the same is attained with the proposed GRNN prediction model with higher accuracy.
- It is also concluded that Tamilnadu consists of optimum parameters (rainfall, temperature, and pH value) for paddy cultivation that enable the farmers to attain higher yield.
- The recommended machine learning algorithm, notably GRNN, reduces the risk factor for paddy yield due its superior performance metrics.

In the future, superlative hybridization among the four adapted machine learning methods will be carried out using additional model performance metrics.

Author Contributions: Conceptualization, V.J. and S.M.P.; methodology, V.J.; software, V.J.; validation, S.M.P. and R.K.; writing—original draft preparation, V.J., S.M.P. and R.K.; writing—review and editing, V.J., S.M.P. and R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Feng, P.; Wang, B.; Liu, D.L.; Waters, C.; Yu, Q. Incorporating machine learning with biophysical model can improve the evaluation of climate extremes impacts on wheat yield in south-eastern Australia. *Agric. For. Meteorol.* **2019**, *275*, 100–113. [CrossRef]
2. Frei, U.; Sporri, S.; Stebler, O.; Holecz, F. Rice field mapping in Sri Lanka using ERS SAR data. *Earth Observ. Q.* **1999**, *63*, 30–35.
3. Rashid, M.; Bari, B.S.; Yusup, Y.; Kamaruddin, M.A.; Khan, N. A Comprehensive Review of Crop Yield Prediction Using Machine Learning Approaches with Special Emphasis on Palm Oil Yield Prediction. *IEEE Access* **2021**, *9*, 63406–63439. [CrossRef]
4. Schlenker, W.; Roberts, M.J. Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 15594–15598. [CrossRef] [PubMed]
5. Folberth, C.; Baklanov, A.; Balkovič, J.; Skalský, R.; Khabarov, N.; Obersteiner, M. Spatio-temporal downscaling of gridded crop model yield estimates based on machine learning. *Agric. For. Meteorol.* **2019**, *264*, 1–15. [CrossRef]
6. Innes, P.J.; Tan, D.K.Y.; Ogtrop, F.V.; Amthor, J.S. Effects of high-temperature episodes on wheat yields in New South Wales, Australia. *Agric. For. Meteorol.* **2015**, *208*, 95–107. [CrossRef]
7. Lamb, D.W.; Brown, R.B. PA—Precision agriculture. *J. Agric. Eng. Res.* **2001**, *78*, 117–125. [CrossRef]
8. Chlingaryan, A.; Sukkarieh, S.; Whelan, B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. *Comput. Electron. Agric.* **2018**, *151*, 61–69. [CrossRef]

9. Mishra, S.; Mishra, D.; Santra, G.H. Applications of Machine Learning Techniques in Agricultural Crop Production: A Review Paper. *Indian J. Sci. Technol.* **2016**, *9*, 1–14. [CrossRef]
10. Marko, O.; Brdar, S.; Panic, M.; Lugonja, P.; Crnojevic, V. Soybean varieties portfolio optimisation based on yield prediction. *Comput. Electron. Agric.* **2016**, *127*, 467–474. [CrossRef]
11. Sellam, V.; Poovammal, E. Prediction of Crop Yield using Regression Analysis. *Indian J. Sci. Technol.* **2016**, *9*, 1–5. [CrossRef]
12. You, J.; Li, X.; Low, M.; Lobell, D.; Ermon, S. Deep Gaussian Process for Crop Yield Prediction Based on Remote Sensing Data. In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17), San Francisco, CA, USA, 4–9 February 2017.
13. Gu, J.; Yin, G.; Huang, P.; Guo, J.; Chen, L. An improved back propagation neural network prediction model for subsurface drip irrigation system. *Comput. Electr. Eng.* **2017**, *60*, 58–65. [CrossRef]
14. Kosari-Moghaddam, A.; Rohani, A.; Kosari-Moghaddam, L.; Esmaeipour-Troujeni, M. Developing a Radial Basis Function Neural Networks to Predict the Working Days for Tillage Operation in Crop Production. *Int. J. Agric. Manag. Dev.* **2018**, *9*, 119–133.
15. Khosla, E.; Dharavath, R.; Priya, R. Crop yield prediction using aggregated rainfall-based modular artificial neural networks and Support vector regression. *Environ. Dev. Sustain.* **2019**, *22*, 5687–5708. [CrossRef]
16. Esfandiarpour-Boroujeni, I.; Karimi, E.; Shirani, H.; Esmaeilizadeh, M.; Mosleh, Z. Yield prediction of apricot using a hybrid particle swarm optimization-imperialist competitive algorithm- support vector regression (PSO-ICA-SVR) method. *Sci. Hortic.* **2019**, *257*, 108756. [CrossRef]
17. Maya Gopal, P.S.; Bhargavi, R. Performance Evaluation of Best Feature Subsets for Crop Yield Prediction Using Machine Learning Algorithms. *Appl. Artif. Intell.* **2019**, *33*, 621–642. [CrossRef]
18. Kim, N.; Ha, K.J.; Park, N.W.; Cho, J.; Hong, S.; Lee, Y.W. A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015. *ISPRS Int. J. Geo-Inf.* **2019**, *8*, 240. [CrossRef]
19. Khaki, S.; Wang, L. Crop Yield Prediction Using Deep Neural Networks. *Front. Plant Sci.* **2019**, *10*, 621. [CrossRef]
20. Abdipour, M.; Younessi-Hmazekhanlu, M.; Ramazani, S.H.R.; Omidi, A.H. Artificial neural networks and multiple linear regression as potential methods for modeling seed yield of safflower (*Carthamus tinctorius* L.). *Ind. Crop. Prod.* **2019**, *127*, 185–194. [CrossRef]
21. Palanivel, K.; Surianarayanan, C. An Approach for Prediction Of Crop Yield Using Machine Learning And Big Data Techniques. *Int. J. Comput. Eng. Technol.* **2019**, *10*, 110–118. [CrossRef]
22. Cai, Y.; Guan, K.; Lobell, D.; Potgieter, A.B.; Wang, S.; Peng, J.; Xu, T.; Asseng, S.; Zhang, Y.; You, L.; et al. Integrating satellite and climate data to predict wheat yield in Australia using machine learning approaches. *Agric. For. Meteorol.* **2019**, *274*, 144–159. [CrossRef]
23. Wang, L.; Wang, P.; Liang, S.; Zhu, Y.; Khan, J.; Fang, S. Monitoring maize growth on the North China Plain using a hybrid genetic algorithm-based back-propagation neural network model. *Comput. Electron. Agric.* **2020**, *170*, 105238. [CrossRef]
24. Abbas, F.; Afzaal, H.; Farooque, A.A.; Tang, S. Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms. *Agronomy* **2020**, *10*, 1046. [CrossRef]
25. Xiong, Y.; Ohashi, S.; Nakano, K.; Jiang, W.; Takizawa, K.; Iijima, K.; Maniwara, P. Application of the radial basis function neural networks to improve the nondestructive Vis/NIR spectrophotometric analysis of potassium in fresh lettuces. *J. Food Eng.* **2021**, *298*, 110417. [CrossRef]
26. Available online: https://mausam.imd.gov.in/ (accessed on 5 August 2021).
27. Available online: https://www.tnagrisnet.tn.gov.in/ (accessed on 5 August 2021).
28. Available online: https://tn.data.gov.in/statedepartment/department-economics-and-statistics (accessed on 5 August 2021).
29. Piekutowska, M.; Niedbała, G.; Piskier, T.; Lenartowicz, T.; Pilarski, K.; Wojciechowski, T.; Pilarska, A.A.; Czechowska-Kosacka, A. The Application of Multiple Linear Regression and Artificial Neural Network Models for Yield Prediction of Very Early Potato Cultivars before Harvest. *Agronomy* **2021**, *11*, 885. [CrossRef]
30. Pandey, A.; Mishra, A. Application of artificial neural networks in yield prediction of potato crop. *Russ. Agric. Sci.* **2017**, *43*, 266–272. [CrossRef]
31. Son, N.T.; Chen, C.F.; Chen, C.R.; Guo, H.Y.; Cheng, Y.S.; Chen, S.L.; Lin, H.S.; Chen, S.H. Machine learning approaches for rice crop yield predictions using time-series satellite data in Taiwan. *Int. J. Remote. Sens.* **2020**, *41*, 7868–7888. [CrossRef]
32. Granata, F. Evapotranspiration evaluation models based on machine learning algorithms—A comparative study. *Agric. Water Manag.* **2019**, *217*, 303–315. [CrossRef]
33. Han, L.; Yang, G.; Dai, H.; Xu, B.; Yang, H.; Feng, H.; Li, Z.; Yang, X. Modeling maize above-ground biomass based on machine learning approaches using UAV remote-sensing data. *Plant Methods* **2019**, *15*, 10. [CrossRef]
34. Ramesh, D.; Vardhan, B.V. Analysis Of Crop Yield Prediction Using Data Mining Techniques. *Int. J. Res. Eng. Technol.* **2015**, *4*, 470–473.
35. Elavarasan, D.; Vincent, P.M.D. Crop yield prediction using deep reinforcement learning model for sustainable agrarian applications. *IEEE Access* **2020**, *8*, 8688686901.
36. Gopal, P.S.M.; Bhargavi, R. A novel approach for efcient crop yield prediction. *Comput. Electron. Agric.* **2019**, *165*, 104968. [CrossRef]
37. Shiu, Y.-S.; Chuang, Y.-C. Yield estimation of paddy rice based on satellite imagery: Comparison of global and local regression models. *Remote Sens.* **2019**, *11*, 111. [CrossRef]

38. Guo, Y.; Fu, Y.; Hao, F.; Zhang, X.; Wu, W.; Jin, X.; Bryant, C.R.; Senthilnath, J. Integrated phenology and climate in rice yields prediction using machine learning methods. *Ecol. Indic.* **2021**, *120*, 106935. [CrossRef]
39. Elavarasan, D.; Vincent, P.M.D.R.; Srinivasan, K.; Chang, C.Y. A hybrid CFS lter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. *Agriculture* **2020**, *10*, 400. [CrossRef]
40. Khaki, S.; Wang, L.; Archontoulis, S.V. A CNN-RNN framework for crop yield prediction. *Front. Plant Sci.* **2020**, *10*, 1750. [CrossRef] [PubMed]
41. Yalcin, H. An approximation for a relative crop yield estimate from field images using deep learning. In Proceedings of the 2019 8th International Conference on Agro-Geoinformatics (Agro-Geoinformatics), Istanbul, Turkey, 16–19 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6.
42. Hilal, Y.Y.; Ishak, W.; Yahya, A.; Asha'ari, Z.H. Development of genetic algorithm for optimization of yield models in oil palm production. *Chil. J. Agric. Res.* **2018**, *78*, 228–237. [CrossRef]
43. Nevavuori, P.; Narra, N.; Lipping, T. Crop yield prediction with deep convolutional neural networks. *Comput. Electron. Agric.* **2019**, *163*, 104859. [CrossRef]
44. Government of India. A Status Note on Rice in India, National Food Security Mission. Available online: https://www.nfsm.gov.in (accessed on 5 August 2021).