ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CROP YIELD PREDICTION IN BIG DATA USING MARGALEF KERNEL PERCEPTRON BASED WINNOW BROWN BOOST CLASSIFIER

[1]S. SARITHA, [2]G. ABEL THANGARAJA
[1]Research Scholar, Department of Computer Science
[1]Sri Nehru Maha Vidyalaya College of Arts and Science
Coimbatore, India.
[2] Department of Computer Technology
[2]Sri Krishna Adithya College of Arts and Science
Coimbatore, India.

Email: [1]sarithajayabrabu@gmail.com, [2]abeltraja@gmail.com

Corresponding author: sarithajayabrabu@gmail.com

ABSTRACT

Crop prediction is a very difficult feature obtained by different characteristics like environment, genotype, and their associations. Policy and decision-makers depend on accurate crop yield predictions to ensure timely import and export recommendations to reinforce food security. In agriculture, boosting in machine learning (ML) is utilized to forecast crop yield. Many boosting ML approaches such as classification, prediction, and clustering predict agricultural production. Data mining techniques are a mandatory technique for achieving significant solutions for this issue. To perform the crop prediction based on different weathers in big data analytics is called as Margalef Kernel Perceptron and Winnow Brown Boosting Classification (MKP-WBBC) method. MKP-WBBC method for big data-based crop yield prediction is split into two sections, namely, feature selection and classification. First Margalef Kernel Perceptron-based feature selection is applied to the Crop Yield Prediction dataset to select computationally efficient features even in case of huge voluminous data. Second, with the unique features selected, Winnow Brown Boosting Classification is applied for accurate and precise crop yield prediction. The main contribution of the new crop yield prediction method is the potentiality to produce accurate predictions and reasonable insights in a simultaneous fashion. This was arrived at by the training and learning algorithm to choose the unique features and not only boosts the results other than enhance the cache hit rate to balance prediction accuracy for training data and generalizability to test data. A discussion of the results achieved reveals the productive performance of the MKP-WBBC method to predict crop yield accurately. Furthermore, results indicate the proposed MKP-WBBC can efficiently enhance crop yield prediction performance and analysis the existing methods in different parameters likes, feature selection accuracy, feature selection time, error rate, and air pollution prediction accuracy.

Keywords: *Machine Learning, Boosting, Margalef Kernel Perceptron, Feature Selection, Winnow Brown Boosting, Classification*

1. INTRODUCTION

Two major characteristics influencing food security are crop quantity and crop quality. This is specifically found to be of more significance as far as developing countries are concerned where agriculture is still found a dominant part of the economy. Also, for predicting crop yield, in-depth knowledge of how much sunlight specific plants receive and the amount of water they require. Moreover, the growth of plants also is influenced by some of the other factors, like the temperature recorded, the humidity it requires, and the nature and type of soil.

An Agricultural Production Systems simulator crop model (APSIM crop model) was proposed in [1] to combine the daily high-resolution CubeSat imagery. The designed model employed APSIM with linear regression to simulate LA). The relationship was employed to identify the optimal regression date wherein LAI was utilized in providing yield prediction. But it failed to improve the prediction accuracy by yield prediction approach.

A DLMLP neural network was introduced in [2] with the aim of addressing crop yield forecast-related issues. Here, remotely sensed data was employed for crop yield forecast. Moreover, a machine learning technique was utilized in predicting crop yield in an accurate manner employing soil health parameters. However, the prediction time was not reduced by DLMLP neural network.

The long-term trend was discussed in [3] to reproduction growth and yield aging effects, yield-related characteristics, and illness resistance features for cereal crops with a spectrum of genotypes. Numerous varieties were developed over distinct environmental conditions under two intensity levels. Breeding progress was better for winter barley by winter rye hybrid and minimum winter rye population varieties. However, the computational cost was not reduced.

A multi-sensor method was designed in [4] for drought-induced agricultural impact prediction. The different input data included MODIS NDVI and LST, ESA-CCI Soil Moisture, and CHIRPS rain data. The designed method is processed at the section level in lightly monitored cropland in Argentina. Lasso regression of rank values with outpost data was determined for relative yield anomalies. The designed method was strong for great drought events. However, the computational complexity was not reduced by the multi-sensor method.

The prevailing unpleasant outcomes of agriculture on the biosphere cannot be minimized by employing conventional methods. The interval between environmentally unfavorable influences and the steadily inflating knowledge base is heightening. Hence, crop yield prediction over the past few decades is considered as a crucial research field, as it has a uniformly significant means of reference for managing farm management in planning and pre-crop processes.

In [5], a new maize harvest prediction method based on spatiotemporal data mining was proposed to identify the viable solution; several models were utilized, like, CP-ANNs, neural networks with ReLU, and XGBoost. With these models distinct subsets of independent variables for the corresponding five vegetation periods were also analyzed, therefore contributing to prediction accuracy. However, both accurate and timely crop prediction assessment at a small scale is important concerning food security and harvest management. A multilevel deep learning method integrating Recurrent Neural Network and Convolutional Neural Network for extracting both spatial and temporal features was presented in [6]. Here, corn yield was said to be predicted for the period of 2013 to 2016.

One of the natural evolutions of sustainable agriculture is farm-scale crop yield prediction. With this type of prediction not only results in abundant food without decrease but also does not pollute environmental factors. Though several crop yield productions have been made with only constrained regional-scale predictions. Multi-temporal data was utilized in [7] with help of a deep hybrid neural network model to train this type of information. But, this absolute error involved in minimized to a greater extent. In [8], yet another method employing Bidirectional Long Short-Term Memory and Bidirectional Gated Recurrent Units was integrated to focus on the prediction error.

1.1 Problem statement

To this end, this paper seeks to undergo a research study on the application of Margalef Kernel Perceptron and Winnow Brown Boosting Classification (MKP-WBBC) based crop yield prediction. Hence, the study aims to unearth a comparison between results obtained with this MKP-WBBC method to learn more about their paramount employment in crop yield prediction. Moreover, our work aims to provide a crop yield prediction method with good accuracy, time, and error rate. In this study, a system for the prediction of crop yield by utilizing an advanced boosting model is made. Two baseline predictive learning methods were also designed in this study for comparison with our proposed method.

1.2 Contributions

The key contributions of this study are:

1. To improve the precise and accurate results, Margalef Kernel Perceptron and Winnow Brown Boosting Classification (MKP-WBBC) that by combining the feature selection and classification models.
2. To reduce the feature selection time and overhead, by proposing the Margalef Kernel Perceptron Feature Selection model, irrelevant features are eliminated and relevant unique features are selected.
3. To combine the advantages of the Kernel Perceptron function and the Margalef Similarity Index, which is efficient at

filtering out irrelevant crop yield features from further processing.

4. To minimize the error rate of prediction results, using Winnow Brown Boosting Classification for accurate and precise crop yield prediction that combines the weak classifiers to form a strong classifier.
5. Comparing the performances of two popular prediction methods in crop yield prediction, the practicality and feasibility of the MKP-WBBC method by comparing the metrics in terms of feature selection time, feature selection overhead, error rate, and crop yield prediction accuracy.

1.3 Work Organization

The remaining article is structured as follows: section 2, introduces the work related to crop yield prediction using machine and boosting techniques. Section 3, describes the MKP-WBBC method for crop yield prediction. Section 4, provides the experimental process and gives a predictive evaluation of the experimental results in the form of a discussion in section 5. Finally, concluding remarks are provided in the sixth section.

2. RELATED WORKS

Crop prediction optimization is a pivotal shift to validate secure and stable food production globally. In the human population, the insistence on food heightens is improved. In order to, create a certain ultimatum, crops require improving production and maintaining production costs and area demands to a minimum.

In [9], a performer-based deep learning method was introduced with aim of predicting crop yield utilizing both single nucleotide polymorphisms and weather data, as a result reducing RMSE.

Machine Learning techniques are employed in several areas of applications, ranging from big departmental stores to business establishments for evaluating customer behavior. Amongst them, predicting crop yield is considered one of the most demanding issues as far as precision agriculture is concerned. Therefore, several methods have been designed recently and also validated. Over the past few years, predicting crop yield can evaluate the actual yield in a reasonable fashion, but improved performance within crop yield prediction is though preferable. A review of deep learning methods for horticultural-related crop yield prediction was investigated in [10]. Yet another comprehensive review of machine learning techniques for smart farming with the purpose of monitoring crop quality and assessing yield was designed in [11].

A detailed review of the capability, advantages, and drawbacks of each method and the appropriateness of these methods under several agricultural circumstances were presented in [12]. A systematic Literature Review (SLR) for extracting and synthesizing algorithms and characteristics that have been utilized for crop yield prediction was designed in [13]. Hydro-agrological research, ETo determined by meteorological factors, significant for achieving accurate irrigation and precision agriculture is concerned. ETo based on a single meteorological factor would be advantageous in places where climatic factors are demanding.

A DNN architecture for predicting daily ETo with a single input factor on the basis of feature importance (FI) score using random forest (RF), and extreme gradient boosting (XGBoost) was performed in [14]. With this FI the accuracy factor was improved. However, it failed to focus on the error rate. Over the past few years, DLMLP neural networks have manifested extraordinary advancement in conveying crop yield prediction-related issues. Crop yield prediction accurately employing machine learning is critical as it aids in keeping a path of soil health and also manipulates the comprehensive yield.

In [15], three significant soil health metrics, like, Soil Moisture, Soil Salinity, and SOC were utilized for crop yield prediction. Crop yield prediction comprises enormous voluminous data found to be suitable for data mining methods. This improvement was said to be obtained both in terms of error and accuracy. Random forest [16] was applied for fast inspection of agricultural yield forecast that in turn focused on both the accuracy and error aspect.

A case scenario of rice manufacture in China employing SVM was performed in [17]. With the advent of artificial intelligence and computer science, data mining has acquired an extensive volume of enhancement. Despite the voluminous information addressed the yield accuracy was not concentrated.

In [18], regression techniques were employed for predicting the yields of wheat, maize, and cotton crops. Finally, Root Mean Squared Error (RMSE) involved in crop yield prediction was also measured. In [19], long-term movements for breeding purposes and yield associated characteristics from German category evaluations for five distinct cereal crops with an extensive genotype extent. Moreover, soybean yield analysis in Argentina was presented in [20].

Motivated by the above said research gaps, like the accuracy aspect, the time factor, and error rate, in this work a method called, Margalef Kernel Perceptron and Winnow Brown Boosting Classification (MKP-WBBC) for crop yield prediction concerning big data analytics is designed. A detailed description of the MKP-WBBC method is provided in the following sections.

3. MARGALEF KERNEL PERCEPTRON AND WINNOW BROWN BOOSTING CLASSIFICATION FOR CROP YIELD PREDICTION

As far as the agricultural field is concerned one of the important issues is crop yield prediction. Every single farmer is consistently making an effort to perceive how much yield will get from his anticipation. Also, over the past few years, with the increase in the available data (i.e., big data) yield prediction was measured by examining the farmer's preceding experience on a specific crop.

Agricultural yield fundamentally is influenced by rainfall, pesticides, temperature as well as the forecast of the yield process. Therefore, accurate crop yield prediction with big data analytics yet remains an extensive issue that needs to be addressed for making decisions concerning agricultural risk management. To overcome this proposed work, Margalef Kernel Perceptron and Winnow Brown Boosting Classification (MKP-WBBC) for crop yield prediction with higher accuracy are performed. In figure 1 describes the block diagram of the MKP-WBBC method for crop yield prediction.

Figure 1, illustrates the proposed MKP-WBBC method includes two processes, namely, feature selection and classification. Initially, with the Crop Yield Prediction dataset provided as input, relevant features were selected employing the Margalef Kernel Perceptron Feature Selection model. Here, the Kernel perceptron being a type of perceptron learning utilized kernel function to determine the similarity between features.

Moreover, with the aid of the Margalef Similarity Index based on the relevance factor between the features efficient feature selections were made. Second, with the selected features as input, Winnow Brown Boosting Classification Process was performed to classify the crop yields. Here, Winnow Classifier is considered as the weak classifier.

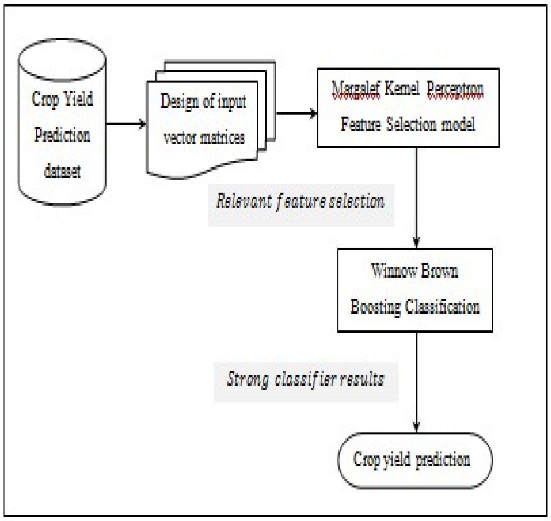

Crop Yield Prediction dataset
Design of input vector matrices
Margalef Kernel Perceptron Feature Selection model
Relevant feature selection
Winnow Brown Boosting Classification
Strong classifier results
Crop yield prediction

Figure 1: Block Diagram Of MKP-WBBC Method

Finally, Brown Boosting Classification is carried out to combine the weak classifiers to form strong classifiers for efficient crop yield prediction. In this way, the crop yield prediction performance gets improved. In the succeeding section, the description of the crop prediction dataset used in the proposed work is initially provided, followed by which an elaborate description of the MKP-WBBC method is presented.

3.1 Crop Yield Prediction Dataset Description

One of the major sectors playing a significant role in the global economy is agriculture. The science of training machines for learning and producing models for future predictions is extensively used over the past few years. Moreover, with the mushrooming widening of the entire human population comprehending global crop yield is key to addressing food security problems and reducing the influences of climate change. Therefore, crop prediction in the recent few years has been determined as a paramount agricultural issue. The agricultural yield also is heavily influenced by several factors like rain, temperature, pesticides, and precise information concerning crop yield history for making decisions concerning agricultural risk management and future predictions.

In whole Crop Prediction Dataset is split into four files namely, pesticide (i.e., pesticide.csv), rainfall (i.e., rainfall.csv), temperature (i.e., temp.csv), yield (i.e., yield.csv), finally integrated into a single file, naming, yield data file (yield_df.csv) respectively. The pesticide file includes features like domain, area, element, item, year, unit, and value. The rainfall file includes area,

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

year, and average rainfall. The temperature file includes the year, country, and average temperature. Finally, the yield file includes the domain code, domain, area code, area, element code, element, item code, item, year code, year, unit, and value. The details are given below.

Pesticide		*Rainfall*		*Temperature*		*Yield*	
S. No	Feature	S. No	Feature	S. No	Feature	S. No	Feature
1	Domain	1	Area	1	Year	1	Domain code
2	Area	2	Year	2	Country	2	Domain
3	Element	3	Average rainfall	3	Average temp	3	Area code
4	Item					4	Area
5	Year					5	Element code
6	Unit					6	Element
7	Value					7	Item code
						8	Item
						9	Year code
						10	Year
						11	Unit
						12	Value

Table 1: Crop Yield Prediction Dataset Description

In above table 1, the Crop Yield Prediction Dataset is provided as input to the proposed method, and four files utilized from it, and four different input vector matrices are given below.

$$P = \begin{bmatrix} P_{11} & P_{12} & \dots & P_{1p} \\ P_{21} & P_{22} & \dots & P_{2p} \\ \dots & \dots & \dots & \dots \\ P_{i1} & P_{i2} & \dots & P_{ip} \end{bmatrix}, where\ i = 7 \quad (1)$$

$$R = \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1r} \\ R_{21} & R_{22} & \dots & R_{2r} \\ \dots & \dots & \dots & \dots \\ R_{j1} & R_{j2} & \dots & R_{jr} \end{bmatrix}, where\ j = 3 \quad (2)$$

$$T = \begin{bmatrix} T_{11} & T_{12} & \dots & T_{1t} \\ T_{21} & T_{22} & \dots & T_{2t} \\ \dots & \dots & \dots & \dots \\ T_{k1} & T_{k2} & \dots & T_{kt} \end{bmatrix}, where\ k = 3 \quad (3)$$

$$Y = \begin{bmatrix} Y_{11} & Y_{12} & \dots & Y_{1y} \\ Y_{21} & Y_{22} & \dots & Y_{2y} \\ \dots & \dots & \dots & \dots \\ Y_{l1} & Y_{l2} & \dots & Y_{ly} \end{bmatrix}, where\ l = 12 \quad (4)$$

The above four matrices represent the input matrices (i.e., obtained from (1), (2), (3) and (4) for pesticide 'P', rainfall 'R', temperature 'T', and yield 'Y' respectively. These four matrices are utilized in the proposed work for further processing (i.e., crop yield prediction).

3.2 Margalef Kernel Perceptron-based Feature selection model

The feature selection is referring to procedure of selecting a subset of important features from a given set of features in a collection of four different input vector matrices (as given in the above section) by dropping irrelevant and redundant features. With this, the computational capacity is said to be improved by decreasing the storage space and therefore improving learning accuracy. Therefore, in feature selection, the principal matter in question lies in efficient means for inspecting the absolute subsets and then examining the absolutely produced subsets. In this work, Margalef Kernel Perceptron-based Feature selection model is employed in inspecting the absolute subsets and then examining the absolutely produced subsets to generate computationally-efficient features. Figure 2 shows the structure of the Margalef Kernel Perceptron based Feature selection model.

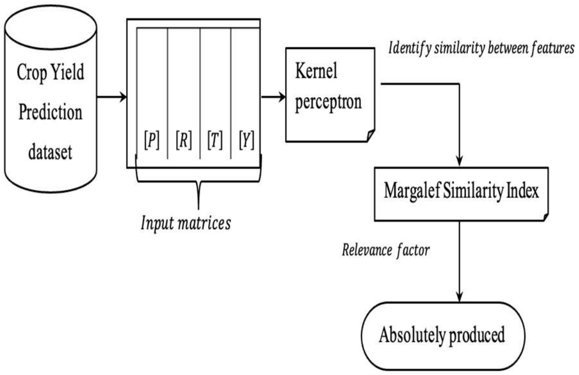
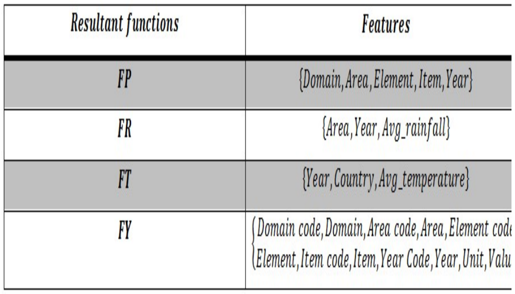

Crop Yield Prediction dataset
[P]
[R]
[T]
[Y]
Input matrices
Kernel perceptron
Identify similarity between features
Margalef Similarity Index
Relevance factor
Absolutely produced

Figure 2: Structure Of Margalef Kernel Perceptron-Based Feature Selection Model

In above figure 2, with the Crop Yield Prediction dataset provided as input and the above four input vector matrices (obtained from (1), (2), (3), and (4)), relevant features are selected employing Margalef Kernel Perceptron Feature Selection model. Here, the Kernel perceptron being a type of perceptron learning utilizes kernel function to determine the similarity between features. Moreover, with the aid of the Margalef Similarity Index based on the relevance factor between the features efficient feature selections were made. Initially, to start with the feature selection process, the features in each input vector matrix have to be obtained. The feature selection process is mathematically stated as,

$$FP \rightarrow ExtractRow[P] \quad (5)$$
$$FR \rightarrow ExtractRow[R] \quad (6)$$
$$FT \rightarrow ExtractRow[T] \quad (7)$$
$$FY \rightarrow ExtractRow[Y] \quad (8)$$

From the above equations (5), (6), (7), and (8) results, the corresponding features in the pesticides file, rainfall file, temperature file and yield file are stored in the respective functions ‘’, ‘’, ‘’, and ‘’ respectively. Then, the features present are stored in the corresponding functions and are given in table 2.



Resultant functions	Features
FP	{Domain,Area,Element,Item,Year}
FR	{Area,Year,Avg_rainfall}
FT	{Year,Country,Avg_temperature}
FY	{Domain code,Domain,Area code,Area,Element code Element,Item code,Item,Year Code,Year,Unit,Valu

Table 2: Resultant Functions And Their Corresponding Features

As obtained from the above table resultant functions and their corresponding features, certain features are said to be repeated and hence those features have to be discarded and only unique features have to be taken for further processing. In our work, the Kernel Perceptron function is applied to identify similarities between features and discard them accordingly.

Let us consider a set of training data as given below.

$$D = \{(A_1, B_1), (A_2, B_2), \ldots . (A_m, B_m)\} \quad (9)$$

$$A_i = \{FP_i \cup FR_i \cup FT_i \cup FY_i\}, B_i \in \{-1, +1\} \quad (10)$$

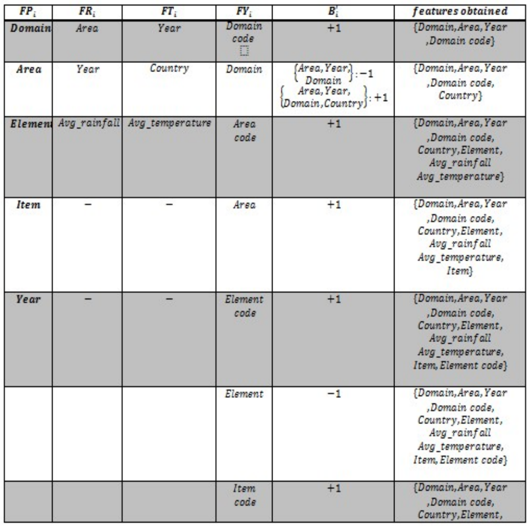
From the above equations (9) and (10), the training data ‘D’ represents the four features that has been stored as a separate function and ‘B_i’ denotes either similar features $(i.e., -1, to\ be\ discarded)$ or dissimilar features $(i.e., +1, to\ be\ included)$ respectively. Here, two functions namely, predict and update is performed. The predict function and the update function is mathematically represented as given below.

$$Predict: B_i' = Sign\ (A_i) \quad (11)$$

$$Predict: B_i' = Sign\ (FP_i, FR_i);\ Sign\ (FP_i, FT_i); Sign \quad (12)$$

$$Update: if\ (\ 〖FP〗\ \_i =\ 〖FR〗\ \_i\ ), result = -1(discard\ feature)\ else\ result = +1(include\ feature) \quad (13)$$

First, from the above three equations (11), (12), and (13), the predict function predicts the presence or absence of similar features and returns ‘-1’ upon identification of similar features or returns ‘$+1$’ upon identification of dissimilar features by employing the update function and accordingly results are obtained as given in the below table 3 refer at the end of the article.



FP_i	FR_i	FT_i	FY_i	B_i'	features obtained
Domain	Area	Year	Domain code □	+1	{Domain,Area,Year ,Domain code}
Area	Year	Country	Domain	$\begin{Bmatrix} Area, Year, \\ Domain \end{Bmatrix}: -1$ $\begin{Bmatrix} Area, Year, \\ Domain, Country \end{Bmatrix}: +1$	{Domain,Area,Year ,Domain code, Country}
Element	Avg_rainfall	Avg_temperature	Area code	+1	{Domain,Area,Year ,Domain code, Country,Element, Avg_rainfall Avg_temperature}
Item	–	–	Area	+1	{Domain,Area,Year ,Domain code, Country,Element, Avg_rainfall Avg_temperature, Item}
Year	–	–	Element code	+1	{Domain,Area,Year ,Domain code, Country,Element, Avg_rainfall Avg_temperature, Item, Element code}
			Element	−1	{Domain,Area,Year ,Domain code, Country,Element, Avg_rainfall Avg_temperature, Item, Element code}
			Item code	+1	{Domain,Area,Year ,Domain code, Country,Element,

Table 3: Kernel Perceptron-Based Feature Similarity Identification Results

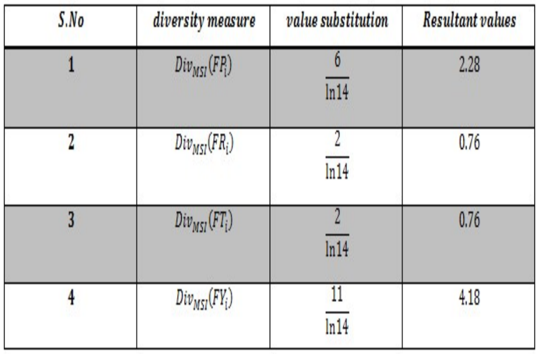
With the above-obtained results as given in table 3, based on the relevance factor employing Margalef Similarity Index feature subset is made. For inspecting the absolute subsets objective measure, i.e., richness indices have been developed to estimate biodiversity (absolutely produced subsets) from field observations (overall samples).

The richness indices ‘RI’ is a measure for total number of the features in a community (i.e., for crop yield). The ‘RI’ is the simplest measure of absolute subsets and is simply a count of the number of different features in a given area (i.e., overall samples). This richness indices 'RI' measure is heavily dependent on sampling size and endeavor and is mathematically formulated as given below.

$$〖Div〗\ \_MSI = (UF - 1)/\ln N \tag{14}$$

From the above equation (14), the diversity measure 'Div_{MSI}' is obtained based on the number of unique features recorded 'UF' and total number of individual features in the sample. Based on the richness indices 'RI' measure, the resultant values obtained are given in table 4.

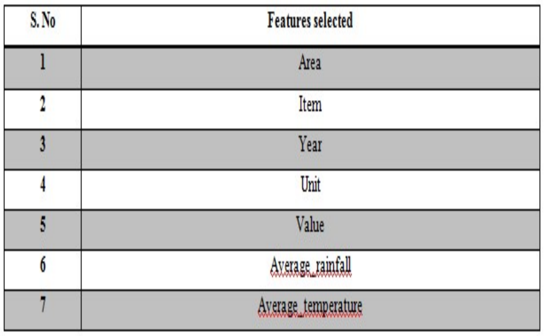
Table 4: Margalef Similarity Index Results



S.No	diversity measure	value substitution	Resultant values
1	$Div_{MSI}(FP_i)$	$\frac{6}{\ln 14}$	2.28
2	$Div_{MSI}(FR_i)$	$\frac{2}{\ln 14}$	0.76
3	$Div_{MSI}(FT_i)$	$\frac{2}{\ln 14}$	0.76
4	$Div_{MSI}(FY_i)$	$\frac{11}{\ln 14}$	4.18

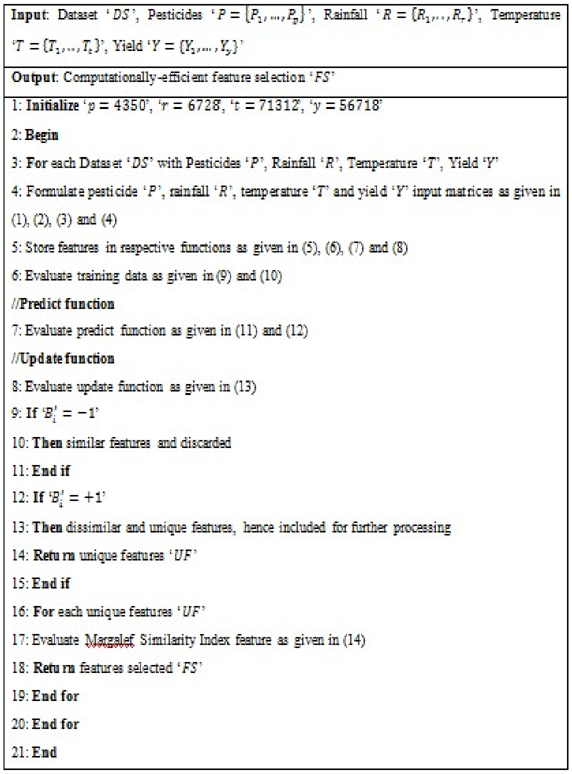
From the above formulations, the features in 'FP_i' should appear '$2\ times$', the features in 'FR_i' should appear '$1\ time$', the features in 'FT_i' should appear '$1\ time$' and the features in 'FY_i' should appear '$4\ times$' respectively. From 'FP_i', year and item occur two times, from 'FR_i' the feature appeared only once is average_rainfall, from 'FT_i' the feature appeared only once is average_temperature and from 'FY_i' no feature appeared four times. With this the final features selected are listed in table 5 given below.

Table 5: Unique Features Selected



S.No	Features selected
1	Area
2	Item
3	Year
4	Unit
5	Value
6	Average_rainfall
7	Average_temperature

On the basis of the above formulated results, the pseudo code representation of Margalef Kernel Perceptron-based feature selection is given below.



```
Input: Dataset 'DS', Pesticides 'P = {P_1, …, P_p}', Rainfall 'R = {R_1, …, R_r}', Temperature 'T = {T_1, …, T_t}', Yield 'Y = {Y_1, …, Y_y}'
Output: Computationally-efficient feature selection 'FS'
1: Initialize 'p = 4350', 'r = 6728', 't = 71312', 'y = 56718'
2: Begin
3: For each Dataset 'DS' with Pesticides 'P', Rainfall 'R', Temperature 'T', Yield 'Y'
4: Formulate pesticide 'P', rainfall 'R', temperature 'T' and yield 'Y' input matrices as given in (1), (2), (3) and (4)
5: Store features in respective functions as given in (5), (6), (7) and (8)
6: Evaluate training data as given in (9) and (10)
//Predict function
7: Evaluate predict function as given in (11) and (12)
//Update function
8: Evaluate update function as given in (13)
9: If '$B_i^t = -1$'
10: Then similar features and discarded
11: End if
12: If '$B_i^t = +1$'
13: Then dissimilar and unique features, hence included for further processing
14: Return unique features 'UF'
15: End if
16: For each unique features 'UF'
17: Evaluate Margalef Similarity Index feature as given in (14)
18: Return features selected 'FS'
19: End for
20: End for
21: End
```

Algorithm 1: Margalef Kernel Perceptron-Based Feature Selection (Features Selected: Area, Item, Year, Unit, Avg_Rainfall, Avg_Tempertaure, Value)

As given in the above algorithm with the objective of selecting computationally efficient features even in the case of big data, first, four distinct matrices are formulated according to the dataset fetched. Second, the similarity of features is obtained by means of Kernel perceptron via two distinct functions, predict and update. Third, with the obtained similar features between four distinct matrices relevance factor is evaluated using the Margalef Similarity Index. Finally, computationally efficient features for further processing are selected.

3.3 Winnow Brown Boosting Classification for Crop Yield Prediction

Boosting is a comprehensively employed model for caching. It improves the model's performance by constructing learners and by focusing on the prediction of crop yields that were rroneously or improperly evaluated by earlier weak learners. Owing to the reason that upcoming weak learners concentrate on the prediction of crop yield on which previous weak learners made errors, the execution of boosting brings about stronger prediction power by reducing bias predominantly. In this work, Winnow Brown Boosting Classification is carried out to integrate the results of weak classifiers for arriving at strong classification results. Figure 3 shows the block diagram of the Winnow Brown Boosting Classification model.

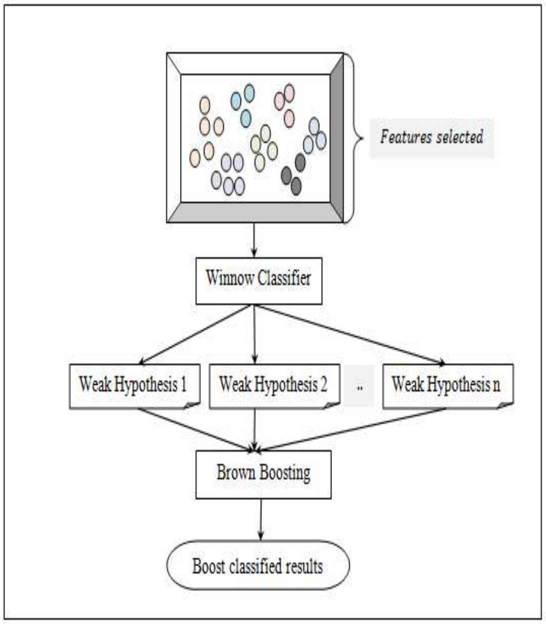


Figure 3: Block Diagram Of Winnow Brown Boosting Classification Model

As shown in the above figure 3, with the features selected results provided as input, the objective of the Winnow Brown Boosting Classification model remains in boosting the features selected by integrating the weak hypothesis (i.e., Winnow Classifier). At this juncture, a Brown Boosting function via normalization factor is introduced that in turn not only boosts the results but also improves cache hit rate by improving the crop yield prediction accuracy.

The main objective of the Winnow Brown Boosting Classification algorithm remains in enhancing the strong classifier's performance by optimally integrating weak classifiers.

The instance space in Winnow Classifier is ‘$A = \{0,1\}$’, that also maintains non-negative weights ‘W_i’ for ‘$i \in (1,2,\dots,n)$’, then, the prediction rule for classification is given below.

$$if\ \sum_{i=1}^{n} W_i A_i > \delta, where\delta = \frac{n}{2} \qquad (15)$$

With the above resultant values, if prediction correctly made, no update is performed. If sample instance is predicted incorrectly where the correct result was ‘0’, for each feature selected ‘FS_i’, the corresponding weight ‘W_i’ is set to ‘0’. On the other hand, if sample instance is predicted incorrectly and the correct result was ‘1’, for each feature selected ‘FS_i’, the corresponding weight ‘W_i’ is multiplied by ‘α’. With the above constructed Winnow Classifier error prediction ‘ε_i’ is made from a weak classifier ‘FS_i’ results from the features selected as given below.

$$\varepsilon_i = \sum_{i=1}^{n} W_i . Ind\ [h_i(FS_i)] \qquad (16)$$

From the above equation (16), ‘$Ind[.]$’ represents the index function that generates the output as ‘1’ if the results of innermost expression is true and ‘0’ if the results of innermost expression is false. In addition, a weight ‘W’ is initially set to ‘1’ to each classified features selected results. Followed by the construction of error function, update function is generated using Brown Boost as given below.

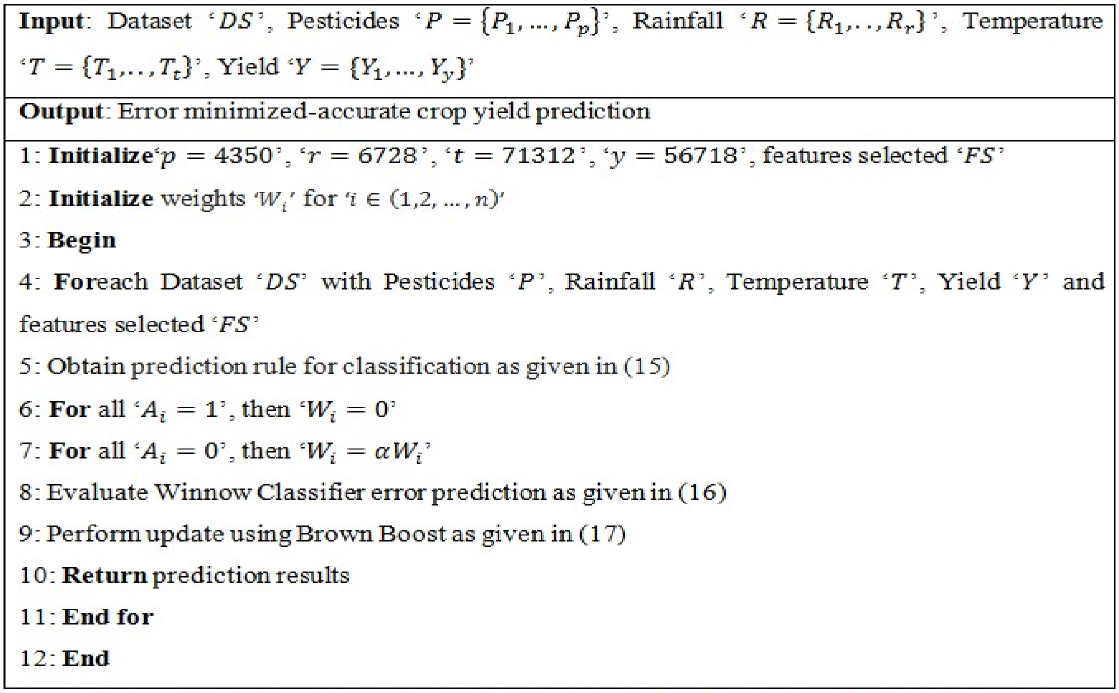
$$PR = r_{i+1}(A_i) = r_i(A_i) + \alpha h_i(FS_i)B_i \qquad (17)$$

From the above equation (17), the value of ‘$r_i(A_i)$’, represent the margin at iteration ‘i’ for example ‘UF_i’. With this the crop yield predicted results are obtained in an accurate fashion. Table 6 given below list the crop yield prediction made for the year between 2012 and 2013 relating to maize yield.

Table 6: Crop Yield Prediction For The Year Between 2012 And 2013 (Concerning Maize)

S. No	Area	Item	Year	Yield	Average_rainfall	Pesticide_tonnes	Average_temperature
91	Albania	Maize	2012	67290	1485	766.25	16.7
95	Albania	Maize	2013	69533	1485	982.32	17.41
203	Algeria	Maize	2012	25583	89	17379.76	17.91
208	Algeria	Maize	2013	33649	89	17278.65	17.65
362	Angola	Maize	2012	7770	1010	40	24.24
370	Angola	Maize	2013	9467	1010	40	24.55
715	Argentina	Maize	2012	57346	591	136185.1	18.18
716	Argentina	Maize	2012	57346	591	136185.1	18.43
731	Argentina	Maize	2013	66037	591	171945.5	16.45
732	Argentina	Maize	2013	66037	591	171945.5	16.88
802	Armenia	Maize	2012	62215	562	278.72	10.2

The table 6 lists the maize crop yield prediction made between the years 2012 and 2013 for 8000 instances. The pseudo-code representation of Winnow Brown Boosting Classification for crop yield prediction is given.



```
Input: Dataset ‘$DS$’, Pesticides ‘$P = \{P_1, \dots, P_p\}$’, Rainfall ‘$R = \{R_1, .., R_r\}$’, Temperature ‘$T = \{T_1, .., T_t\}$’, Yield ‘$Y = \{Y_1, \dots, Y_y\}$’
Output: Error minimized-accurate crop yield prediction
Initialize ‘$p = 4350$’, ‘$r = 6728$’, ‘$t = 71312$’, ‘$y = 56718$’, features selected ‘$FS$’
Initialize weights ‘$W_i$’ for ‘$i \in (1,2, \dots, n)$’
Begin
Foreach Dataset ‘$DS$’ with Pesticides ‘$P$’, Rainfall ‘$R$’, Temperature ‘$T$’, Yield ‘$Y$’ and features selected ‘$FS$’
    Obtain prediction rule for classification as given in (15)
    For all ‘$A_i = 1$’, then ‘$W_i = 0$’
    For all ‘$A_i = 0$’, then ‘$W_i = \alpha W_i$’
    Evaluate Winnow Classifier error prediction as given in (16)
    Perform update using Brown Boost as given in (17)
    Return prediction results
End for
End
```

Algorithm 2 Winnow Brown Boosting Classification for crop yield prediction

In above algorithm 2, with the objective of predicting the correct types of crop yield for the corresponding area, Winnow Brown Boosting Classification is designed. Here, with the unique features selected as input is passed on to the classification process wherein two distinct procedures, namely, prediction and update are performed. First, prediction is done by employing Winnow Classifier. Followed by which in case of erroneous prediction, the Brown Boosting model is applied to reinforce the classification process, therefore resulting in the improvement of overall crop yield prediction accuracy.

4. EXPERIMENTAL SETUP

The performance of the proposed MKP-WBBC method for performing crop yield prediction based on climatic variables in big data analytics several experiments are conducted to show its efficiency through a comparative study including powerful boosting methods, namely APSIM crop model [1] and Deep Learning Multi-Layer Perceptron (DLMLP) [2]. We also compared the performance of existing methods [1] and [2] in Python language using the Crop Yield Prediction dataset(https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset) to measure the efficiency in terms of feature selection accuracy, feature selection time, error rate and crop yield prediction accuracy with respect to a number of features and data points.

5. RESULT AND DISCUSSION

In this section, the results analysis of four distinct performances of improved feature selection accuracy, minimum feature selection time, lesser error rate, and enhanced crop yield prediction accuracy are provided using table and graphical representations.

5.1 Feature selection overhead

The first and foremost performance metric used for crop yield prediction is feature selection overhead. During the selection of unique features, the intermittent values generated are stored in the cache for further processing. This results in a small utilization of overhead and is referred to as the feature selection overhead. The feature selection overhead is obtained as given below.

$$FSO = \sum_{i=1}^{n} Samples_i * Mem\,[FS] \qquad (18)$$

In equation (18), feature selection overhead 'FSO' is measured based on overall samples present in the crop yield prediction dataset '$sampels_i$' and the memory consumed in storing the intermittent process '$Mem[FS]$'. It is measured in terms of kilobytes (KB).

Table 7 Feature Selections Overhead Performance Comparison of the Proposed MKP-WBBC Method Using Crop Yield Prediction Dataset

Samples	Feature selection overhead (KB)		
	MKP-WBBC	APSIM crop model	DLMLP
5000	2750	3500	4250
10000	3125	3735	4815
15000	3350	4155	5325
20000	3825	4835	5855
25000	4135	5215	6035
30000	4255	5815	6385
35000	4835	6135	7455
40000	5125	6245	7835
45000	5835	6835	8135
50000	6015	7135	8935

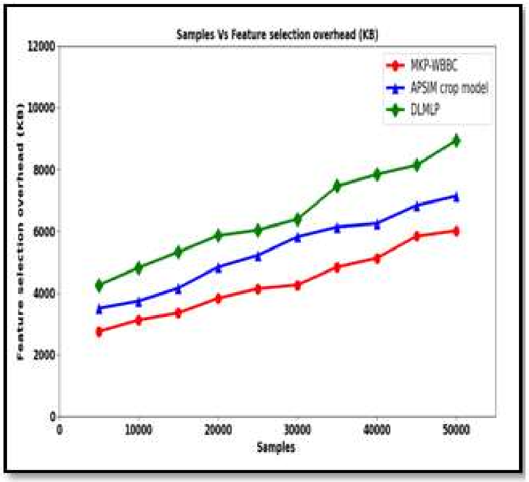


Figure 4: Performance Comparison Of Feature Selection Overhead

Figure 4 illustrates the graphical representation of feature selection overhead for 50000 distinct samples. With the presence of four distinct data files in the crop yield prediction dataset, certain features are found to be repeated and also certain other features are of less significance. Hence, there remains an objective in selecting unique features. So, while performing this process, intermittent features are stored in the stack, resulting in consuming a small portion of memory. From the above figure, increasing the sample cases also causes an increase in the overhead. However, in simulations performed with 5000 samples, 2750KB was utilized by employing the MKP-WBBC method, 3500KB was utilized by employing [1] and 4250KB was utilized by employing [2]. As a result, the overhead incurred using the MKP-WBBC method was found to be comparatively minimum [1] and [2]. The reason behind the improvement was due to the incorporation of the Margalef Kernel Perceptron-based feature selection algorithm. With the applying, this algorithm similarity of features was initially obtained using Kernel perceptron by means of two distinct functions, predict and update. Next, only with the obtained similar features further processing of selecting unique features was made based on the relevance factor employing Margalef Similarity Index.

As a result, during the unique feature selection process, not all the features were utilized, therefore the MKP-WBBC method used minimum overhead by 20% and 34% compared to [1] and [2].

5.2 Feature selection time

One of the most significant performance metrics for crop yield prediction is the time consumed in feature selection. The faster and better the feature selection is made, the further efficient the method is said to be. Feature selection time is measured as,

$$FST = \sum_{i=1}^{n} F_i * Time[UF] \qquad (19)$$

In the above equation (19), ‘is feature selection time measured on basis of the number of features considered for simulation purpose ‘’ and the actual time consumed in obtaining unique features ‘. Feature selection time is measured in terms of milliseconds (ms). Table 8 illustrates the feature selection time comparative results of the considered methods, MKP-WBBC, APSIM crop model [1], and DLMLP [2] respectively.

Table 8 Feature selection time performance comparison of the proposed MKP-WBBC method using crop yield prediction dataset

Samples	Feature selection time (ms)		
	MKP-WBBC	APSIM crop model	DLMLP
5000	125	200	235
10000	145	225	270
15000	175	285	335
20000	215	315	385
25000	285	335	435
30000	315	385	495
35000	335	425	515
40000	350	485	535
45000	385	525	615
50000	415	585	650

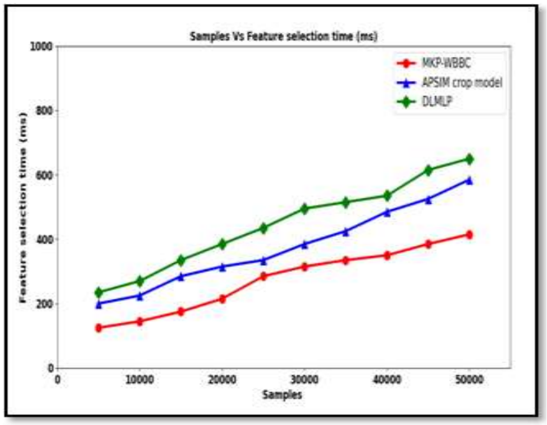


Figure 5: Performance Comparison Of Feature Selection Time

To study the influence of feature selection time on the distinct numbers of crop yield data ranging between 5000 and 50000, experiments were performed with different thresholds (i.e., pesticides, rainfall, temperature, and finally yield basis). The trends of the feature selection time involved with the above said different thresholds can be seen in figure 5. We note that as the number of samples is increased, the percentage of maximum perceptrons is decreased the percentage of feature selection time decreases steadily. However, the homogenized crop yield data sample in terms of crop data increases with the increase in the features being analyzed. As result, feature selection time is reduced by using the MKP-WBBC method by 28% and 40%compared to [1] and [2] respectively. Also only based on the identification of similarity between features and relevance factor, unique features are selected that in turn minimizes the overall feature selection time.

5.3 Crop yield prediction accuracy

Prediction accuracy is measured by dividing the number of correct crop yield predictions made by the total number of predictions. The prediction accuracy is measured as,

$$PA = \sum_{i=1}^{n} \frac{P_{correct}}{P_i} * 100 \qquad (20)$$

In above equation (20), ‘PA’ is prediction accuracy. ‘$P_{correct}$’ Is a number of correct predictions made and total number of predictions ‘P_i’ respectively. It is measured in terms of percentage (%). Table 9 shows the crop yield prediction accuracy performance comparison of the three methods, MKP-WBBC, APSIM crop model [1] and DLMLP [2] based on crop yield prediction dataset.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 9 Crop yield prediction accuracy performance comparison of the proposed MKP-WBBC method using crop yield prediction dataset

Samples	Crop yield prediction accuracy (%)		
	MKP-WBBC	APSIM crop model	DLMLP
5000	96.5	95	92.7
10000	94.15	93.25	91.55
15000	93	92	89.15
20000	92.85	91.35	87
25000	92.55	91	86.35
30000	92	89.55	85.25
35000	90	87.15	84.35
40000	89.15	86.55	82
45000	89	84.35	81
50000	87.35	83	81

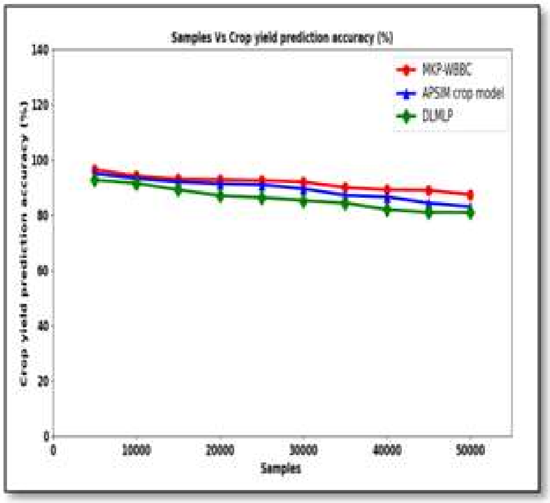


Figure 6: Performance Comparison Of Crop Yield Prediction Accuracy

Figure 6 given above shows the crop yield prediction accuracy was measured using the three methods MKP-WBBC, APSIM crop model [1], and DLMLP [2] taking into consideration 50000 distinct samples. Sample results of ground truth and predicted crop yield prediction accuracy for the case of crop yield are shown in figure 6.

The proposed MKP-WBBC method provides crop yield data being accurately predicted via smooth labeling and it was able to predict multiple classes based on the Brown Boost.

The differences between any pair of observations in the output prediction and ground-truth prediction for each distinct crop yield data ranging between 5000 and 50000 are also provided. On average minimum accuracy differs from 87.35% to 96.5% using MKP-WBBC, 83% to 95% using [1], and 81% to 92.7% using [2] respectively. The MKP-WBBC method was improved in ensuring the crop prediction accuracy level by 3% and 7% compared to [1] and [2]. The reasons behind the improvement using the MKP-WBBC method were selecting the computationally efficient features by means of the Kernel Perceptron function and predict/update function. As a result, with accurate crop yield prediction, decisions can be made in an efficient manner at global, regional, and field levels.

5.4 Error rate

Finally, the error rate or erroneous measure of crop yield prediction made is measured. While predicting crop yield a significant amount of yield is said to be wrongly judged, therefore resulting in the error rate. The error rate is measured as,

$$ER = \sum_{i=1}^{n} \frac{P_{incorrect}}{P_i} \qquad (21)$$

In equation (21), the error rate 'ER' is measured based on the number of incorrect predictions made '$P_{incorrect}$' and the total number of predictions 'P_i' respectively. It is measured in terms of percentage. Finally, table 10 given below lists the error rate.

Table 10 Error rate performance comparison of the proposed MKP-WBBC method using crop yield prediction dataset

Samples	Error rate (%)		
	MKP-WBBC	APSIM crop model	DLMLP
5000	0.7	0.9	1.4
10000	0.8	1.12	1.55
15000	1.05	1.19	1.7
20000	1.18	1.3	2.05
25000	1.25	1.45	2.25
30000	1.3	1.52	2.35
35000	1.35	1.6	2.5
40000	1.42	1.75	2.85
45000	1.49	1.83	3.05
50000	1.53	1.92	3.15

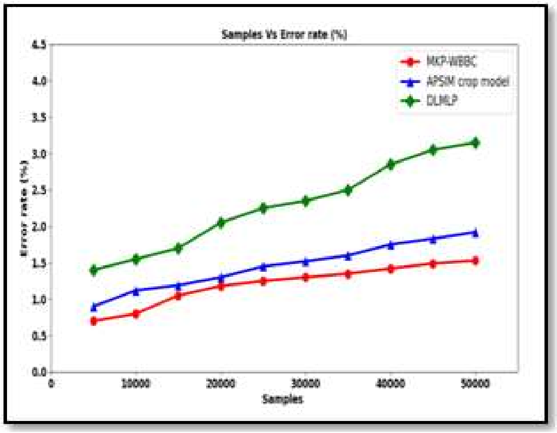


Figure 7: Performance Comparison Of Error Rate

Finally, figure 7 given above illustrates the error rate in the y-axis for distinct numbers of samples ranging between 5000 and 50000 in the x-axis respectively. In the above figure increasing the number of samples in the simulation process results and increase the error rate by using three methods MKP-WBBC, APSIM crop model [1], and DLMLP [2] respectively. However, with simulations performed using 5000 numbers of samples 35 samples were wrongly predicted using MKP-WBBC, 47 numbers of samples were wrongly predicted using [1] and 70 numbers of samples were wrongly predicted using [2], the overall error rate using the three methods was found to be 0.7, 0.9, and 1.4 respectively. As result, the error rate using the MKP-WBBC method was minimizing the existing methods compared to [1] and [2]. The reason behind the improvement was the application of the Winnow Brown Boosting Classification algorithm. Using this algorithm, the first hypothesis was generated for distinct unique features selected. Followed by which erroneous predictions were validated using Winnow Classifier. Finally, the correct predictions were ensemble using Brown Boosting Classification that in turn aided in improving the crop yield prediction to a greater extent. Finally, the error rate of crop yield prediction in the MKP-WBBC method was lesser by 17% and 47% compared to [1] and [2].

6. CONCLUSION

Existing methods to predict crop yield method detect yield by means of soil health parameters and linear regression to simulate leaf area index using linear regression and neural network employ classification models from the non-linear approximation's attributes. The Margalef Kernel Perceptron and Winnow Brown Boosting Classification (MKP-WBBC) methods are capable of obtaining computationally efficient and robust features due to the kernelization of crop yield based on the Margalef Similarity Index using relevance factor. The limitation of minimizing the error rate or falsely predicting area with the corresponding crop yield though not, Winnow Brown Boosting Classification Process is designed with the aid of selected unique features. With this, the MKP-WBBC method precisely and accurately predicts crop yield and also minimizes error rate for further analysis in a timely manner. Experiments on the crop yield prediction dataset describe the MKP-WBBC method as relatively better than the existing method in different parameters likes, like feature selection overhead, feature selection time, error rate, and crop yield prediction accuracy.

Data availability statement
(https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset) for accuracy comparison.

Declaration of interests statement
The authors declare no conflict of interest.

Additional information
No additional information is available for this paper.

REFERENCES:

[1] Matteo G. Ziliani, Muhammad U. Altaf, Bruno Aragon, Rasmus Houborg, Trenton E. Franz, Yang Lu, Justin Sheffield, Ibrahim Hoteit, Matthew F. McCabe, "Early season prediction of within-field crop yield variability by assimilating CubeSat data into a crop model", Agricultural and Forest Meteorology, Elsevier, Volume 313,2022, Pages 1-15.

[2] Akshar Tripathi, Reet Kamal Tiwari and Surya Prakash Tiwari, "A deep learning multi-layer perceptron and remote sensing approach for soilhealth-based crop yield estimation", International Journal of Applied Earth Observations and Geoinformation, Elsevier, Volume 113,2022, Pages 1- 12.

[3] F. Laidig, T. Feike, B. Klocke, J. Macholdt, T. Miedaner, D. Rentel and H. P. Piepho, "Long-term breeding progress of yield, yield-related, and disease resistance traits in five cereal crops of German variety trials", Theoretical and Applied Genetics, Springer, Volume 134, 2021, Pages 3805–3827.

[4] Martin D. Maas, Mercedes Salvia, Pablo C. Spennemann and Maria Elena Fernandez-Long, "Robust Multi sensor Prediction of Drought-Induced Soybean Yield

Anomalies in Argentina", IEEE Geoscience and Remote Sensing Letters, Volume 19, May 2022, Pages 1-6.

[5] A. Nyeki, C. Kerepesi, B. Daroczy, A. Benczur, G. Milics, J. Nagy, E. Harsanyi, A. J. Kovacs, M. Nemenyi, "Application of spatio-temporal data in site-specific maize yield prediction with machine learning methods", Precision Agriculture, Springer, Aug 2021.

[6] Jie Sun, Zulong Lai, Liping Di, Ziheng Sun, Jianbin Tao, and Yonglin Shen, "Multilevel Deep Learning Network for County-LevelCorn Yield Estimation in the U.S. Corn Belt", IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, Vol. 13, 2020.

[7] Martin Engen, Erik Sando, Benjamin Lucas Oscar Sjolander, Simon Arenberg, Rashmi Gupta, Morten Goodwin, "Farm-Scale Crop Yield Prediction from Multi-Temporal Data Using Deep Hybrid Neural Networks", Creative Commons Attribution, Agronomy, Aug 2021.

[8] Khadijeh Alibabaei, Pedro D. Gaspar, Tânia M. Lima, "Crop Yield Estimation Using Deep Learning Based on Climate Big Data and Irrigation Scheduling", Creative Commons Attribution, Energies, Oct 2021.

[9] Hakon Maloy, Susanne Windju, Stein Bergersen, Muath Alsheikh, Keith L. Downing, "Multimodal performers for genomic selection and crop yield prediction", Smart Agricultural Technology, Oct 2021.

[10] Biyun Yang and Yong Xu, "Applications of deep-learning approaches in horticultural research: a review", Horticulture Research, Springer, Aug 2021.

[11] Abhinav Sharma, Arpit Jain, Prateek Gupta, Vinay Chowdary, "Machine Learning Applications for Precision Agriculture: A Comprehensive Review", IEEE Access, Dec 2020.

[12] Abdelraouf M. Ali, Mohamed Abouelghar, A.A. Belal, Nasser Saleh, Mona Yones, Adel I. Selim, Mohamed E.S. Amin, AmanyElwesemy, Dmitry E. Kucher, Schubert Maginan, Igor Savin, "Crop Yield Prediction Using Multi Sensors Remote Sensing (Review Article)", The Egyptian Journal of Remote Sensing and Space Sciences, Elsevier, May 2022.

[13] Thomas van Klompenburg, Ayalew Kassahun, Cagatay Catal, "Crop yield prediction using machine learning: A systematic literature review", Computers and Electronics in Agriculture, Aug 2020.

[14] Sowmya Mangalath Ravindran, Santosh Kumar Moorakkal Bhaskaran, Sooraj Krishnan Nair Ambat, "A Deep Neural Network Architecture to Model Reference Evapotranspiration Using a Single Input Meteorological Parameter", Environmental Processes, Springer, Oct 2021.

[15] Akshar Tripathi, Reet Kamal Tiwari, Surya Prakash Tiwari, "A deep learning multi-layer perceptron and remote sensing approach for soil health based crop yield estimation", International Journal of Applied Earth Observations and Geoinformation, Elsevier, Aug 2022.

[16] Pallavi Kamath, Pallavi Patil, Shrilatha S, Sushma, Sowmya S, "Crop yield forecasting using data mining", Global Transitions Proceedings, Jul 2021.

[17] Su Ying-xue, Xu Huan, Yan Li-jiao, "Support vector machine-based open crop model (SBOCM): Case of rice production in China", Saudi Journal of Biological Sciences, Springer, Jan 2017.

[18] Aditya Shastry, HA Sanjay and E. Bhanusree, "Prediction of Crop Yield Using Regression Techniques", International Journal of Soft Computing, Oct 2017.

[19] F. Laidig, T. Feike, B. Klocke, J. Macholdt, T. Miedaner, D. Rentel, H. P. Piepho, "Long-term breeding progress of yield, yield-related, and disease resistance traits in five cereal crops of German variety trials", Theoretical and Applied Genetics, Springer, Oct 2021.

[20] Maas, Martin D, Salvia, Mercedes, Spennemann, Pablo, Fernandez-Long, María Elena, "Robust Multisensor Prediction of Drought-Induced SoybeanYield Anomalies in Argentina", IEEE Access, Oct 2021.

Appendix

Table 3: Kernel Perceptron-Based Feature Similarity Identification Results

FP_i	FR_i	FT_i	FY_i	B_i'	features obtained
Domain	Area	Year	Domain code	$+1$	{Domain, Area, Year, Domain code}
Area	Year	Country	Domain	$\begin{Bmatrix}Area, Year,\\ Domain\end{Bmatrix}: -1$ $\begin{Bmatrix}Area, Year,\\ Domain, Country\end{Bmatrix}: +1$	{Domain, Area, Year, Domain code, Country}
Element	Avg_rainfall	Avg_temperature	Area code	$+1$	{Domain, Area, Year, Domain code, Country, Element, Avg_rainfall Avg_temperature}
Item	–	–	Area	$+1$	{Domain, Area, Year, Domain code, Country, Element, Avg_rainfall Avg_temperature, Item}
Year	–	–	Element code	$+1$	{Domain, Area, Year, Domain code, Country, Element, Avg_rainfall Avg_temperature, Item, Element code}
			Element	-1	{Domain, Area, Year, Domain code, Country, Element, Avg_rainfall Avg_temperature, Item, Element code}
			Item code	$+1$	{Domain, Area, Year, Domain code, Country, Element, Avg_rainfall Avg_temperature, Item, Element code, Item code}
			Item	-1	{Domain, Area, Year, Domain code, Country, Element, Avg_rainfall Avg_temperature, Item, Element code, Item code}

			$Year\ Code$	$+1$	$\{Domain, Area, Year, Domain\ code, Country, Element, Avg\_rainfall\ Avg\_temperature, Item, Element\ code, Item\ code, Year\ code\}$
			$Year$	-1	$\{Domain, Area, Year, Domain\ code, Country, Element, Avg\_rainfall\ Avg\_temperature, Item, Element\ code, Item\ code, Year\ code\}$
			$Unit$	$+1$	$\{Domain, Area, Year, Domain\ code, Country, Element, Avg\_rainfall\ Avg\_temperature, Item, Element\ code, Item\ code, Year\ code, Unit\}$
			$Value$	$+1$	$\{Domain, Area, Year, Domain\ code, Country, Element, Avg\_rainfall\ Avg\_temperature, Item, Element\ code, Item\ code, Year\ code, Unit, Value\}$

```
Input: Dataset '$DS$', Pesticides '$P = \{P_1, \dots, P_p\}$', Rainfall '$R = \{R_1, .., R_r\}$', Temperature '$T = \{T_1, .., T_t\}$', Yield '$Y = \{Y_1, \dots, Y_y\}$'
Output: Computationally-efficient feature selection '$FS$'
Initialize '$p = 4350$', '$r = 6728$', '$t = 71312$', '$y = 56718$'
Begin
For each Dataset '$DS$' with Pesticides '$P$', Rainfall '$R$', Temperature '$T$', Yield '$Y$'
    Formulate pesticide '$P$', rainfall '$R$', temperature '$T$' and yield '$Y$' input matrices as given in (1), (2), (3) and (4)
    Store features in respective functions as given in (5), (6), (7) and (8)
    Evaluate training data as given in (9) and (10)
    //Predict function
    Evaluate predict function as given in (11) and (12)
    //Update function
    Evaluate update function as given in (13)
    If '$B_i' = -1$'
        Then similar features and discarded
    End if
```

```
If '$B_i' = +1$'
Then dissimilar and unique features, hence included for further processing
Return unique features '$UF$'
End if
For each unique features '$UF$'
Evaluate Margalef Similarity Index feature as given in (14)
Return features selected '$FS$'
End for
End for
End
```

Algorithm 1: Margalef Kernel Perceptron-Based Feature Selection (Features Selected: Area, Item, Year, Unit, Avg_Rainfall, Avg_Tempertaure, Value)

Table 6: Crop Yield Prediction For The Year Between 2012 And 2013 (Concerning Maize)

S. No	Area	Item	Year	Yield	Average_ Rainfall	Pesticide_Tonnes	Average_ Temperature
91	Albania	Maize	2012	67290	1485	766.25	16.7
95	Albania	Maize	2013	69533	1485	982.32	17.41
203	Algeria	Maize	2012	25583	89	17379.76	17.91
208	Algeria	Maize	2013	33649	89	17278.65	17.65
362	Angola	Maize	2012	7770	1010	40	24.24
370	Angola	Maize	2013	9467	1010	40	24.55
715	Argentina	Maize	2012	57346	591	136185.1	18.18
716	Argentina	Maize	2012	57346	591	136185.1	18.43
731	Argentina	Maize	2013	66037	591	171945.5	16.45
732	Argentina	Maize	2013	66037	591	171945.5	16.88
802	Armenia	Maize	2012	62215	562	278.72	10.2
805	Armenia	Maize	2013	67114	562	278.72	11.08
1690	Australia	Maize	2012	64654	534	48687.88	16.84
1691	Australia	Maize	2012	64654	534	48687.88	20.08
1692	Australia	Maize	2012	64654	534	48687.88	11.79
1693	Australia	Maize	2012	64654	534	48687.88	14.27
1694	Australia	Maize	2012	64654	534	48687.88	19.55
1695	Australia	Maize	2012	64654	534	48687.88	17.47
1732	Australia	Maize	2013	64441	534	45177.18	17.39
1733	Australia	Maize	2013	64441	534	45177.18	20.14
1734	Australia	Maize	2013	64441	534	45177.18	12.19
1735	Australia	Maize	2013	64441	534	45177.18	14.74
1736	Australia	Maize	2013	64441	534	45177.18	19.98
1737	Australia	Maize	2013	64441	534	45177.18	18.09
1860	Austria	Maize	2012	107025	1110	3563.3	9.42
1865	Austria	Maize	2013	81173	1110	3108.6	10.03
1974	Azerbaijan	Maize	2012	51029	447	516.35	13
1980	Azerbaijan	Maize	2013	53903	447	489.8	14.12
2046	Bahamas	Maize	2012	73404	1292	268.2	25.46
2050	Bahamas	Maize	2013	74465	1292	268.2	25.88
2348	Bangladesh	Maize	2012	65838	2666	13289.18	26.28
2349	Bangladesh	Maize	2012	65838	2666	13289.18	26.28
2362	Bangladesh	Maize	2013	65953	2666	15330.16	26.59