ISSN Print: 2664-844X
ISSN Online: 2664-8458
NAAS Rating: 4.97
IJAFS 2025; 7(8): 385-400
www.agriculturaljournals.com
Received: 10-05-2025
Accepted: 14-06-2025

Kanwal Preet Singh Attwal
Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

Corresponding Author:
Kanwal Preet Singh Attwal
Department of Computer Science and Engineering, Punjabi University, Patiala, Punjab, India

Integrated machine learning model for wheat yield prediction using agronomic and meteorological factors: A case study from Punjab, India

Kanwal Preet Singh Attwal

DOI: https://www.doi.org/10.33545/2664844X.2025.v7.i8f.636

Abstract
The study aims to develop a Prediction Model for the wheat yield based on meteorological and agronomic factors using Data Mining techniques. The target area of the Research is the cultivators in the Patiala district of Punjab, India. The researcher has applied the knowledge of the wheat morphology and phenology to build the model. Crop yield is affected by several agronomic factors such as soil type and date of sowing, and meteorological factors such as temperature and rainfall. The existing models apply classification or regression techniques to all the identified factors for the prediction of crop yield. This study divides the set of factors into two categories - the factors which are responsible for year-wise variation in yield, and the factors which are responsible for the individual variation of yield for a particular year among various cultivators. It is found that the year-wise yield variation for a particular cultivator (or a particular region) may be attributed to meteorological factors, whereas the agronomic factors are responsible for variation in yield among different cultivators. So, two models have been proposed; the first model - henceforth known as Block-wise Average Yield Prediction model (BAY model) predicts the Block-wise Average Yield based on temperature, rainfall and the yield data, and the second model - henceforth known as Yield Class Prediction model (YC model) predicts the Yield Class based on soil, management practices and yield data. Finally, these models, i.e., BAY model and YC model are integrated into the final model - Final Yield Prediction model (FYP model) to predict the final yield of a particular cultivator.

Keywords: Wheat yield prediction, factors affecting wheat yield, crop prediction model, agronomic factors affecting wheat yield

1. Introduction
Wheat is a staple crop in Punjab, India, particularly in the winter season, contributing significantly to the region's economy and food security. Given the critical importance of wheat yield, predicting its output accurately can help cultivators optimize their farming practices and support governmental planning for food supply and market stabilization. Since an effective and accurate prediction of the wheat yield is essential to control the forward marketing and to maintain the food security, many scientific researches have been conducted in diverse fields. However, the optimal outcomes can be attained through interdisciplinary research in the field of agriculture engineering and computer engineering [1]. Additionally, with the emergence of machine learning, remote sensing, and big data platforms like WEKA, the accuracy and scalability of yield prediction models have improved significantly [2, 3, 4]. A study has been conducted to develop a model for prediction of wheat yield in Patiala district of Punjab, India [5]. As part of this study, a framework has already been proposed to develop a model for crop yield prediction [6]. To develop the wheat yield prediction model, it is pertinent to understand the morphology and the phenology of the wheat plant. A comprehensive study on the morphology of wheat in the context of developing a predictive model for wheat yield has been conducted [7]. A study has also been conducted to understand the phenology of wheat crop and to determine the factors affecting wheat yield [8]. This study uses meteorological data and agronomic variables collected over five years. Similar methods of integrating remote sensing data and ML algorithms have shown promise in comparable agroclimatic settings [9, 10, 11, 12].

Crop yield is affected by several agronomic factors such as soil type and date of sowing, and meteorological factors such as temperature and rainfall [18]. The study divides the set of factors into two categories - the factors which are responsible for year-wise variation in yield, and the factors which are responsible for the individual variation of yield for a particular year among various cultivators. It is found that the year-wise yield variation for a particular cultivator (or a particular region) may be attributed to meteorological factors, whereas the agronomic factors are responsible for variation in yield among different cultivators. So, two models have been proposed; the first model - known as Block-wise Average Yield Prediction model (BAY model) predicts the Block-wise Average Yield based on temperature, rainfall and the yield data, and the second model - known as Yield Class Prediction model (YC model) predicts the Yield Class based on soil, management practices and yield data. Finally, these models i.e. BAY model and YC model are integrated into the final model - Final Yield Prediction model (FYP model) to predict the final yield of a particular cultivator. The BAY model which predicts the Block-wise Average Yield has already been developed [13]. The present work deals with development of YC model to predict the yield class of a particular cultivator based on various agronomic factors and integration of block-wise average yield obtained from BAY model and yield class obtained from YC model into final yield of a particular cultivator using FYP model.

2. Materials and Methods

Recent studies highlight the importance of ensemble learning and explainable AI in agricultural predictions [14]. Ensemble-based models have proven effective in yield prediction using satellite and UAV data [15]. Multi-task deep learning models, such as MT-CYP-Net and explainable Bi-LSTM, have also been validated in similar contexts [16, 17, 18]. These studies support the hybrid approach taken in this research using WEKA and SPSS.

Figure 1 shows the outline for development of Yield Class model (YC model). To develop this model, the researcher examines the effect of agronomic factors, such as soil texture and management practices on the inter-region variation of the wheat yield in the Patiala district of Punjab, India. The soil and management practices data along with the yield data is used to create the YC model. To find the variations in the yield of different cultivators based on the soil in their farms and the management practices they follow, the variations caused by other factors such as temperature and rainfall, and inter-block variation should be kept as constant. The yield of each cultivator is assigned to a yield-class based on the average yield (AY) of the block for that particular year.

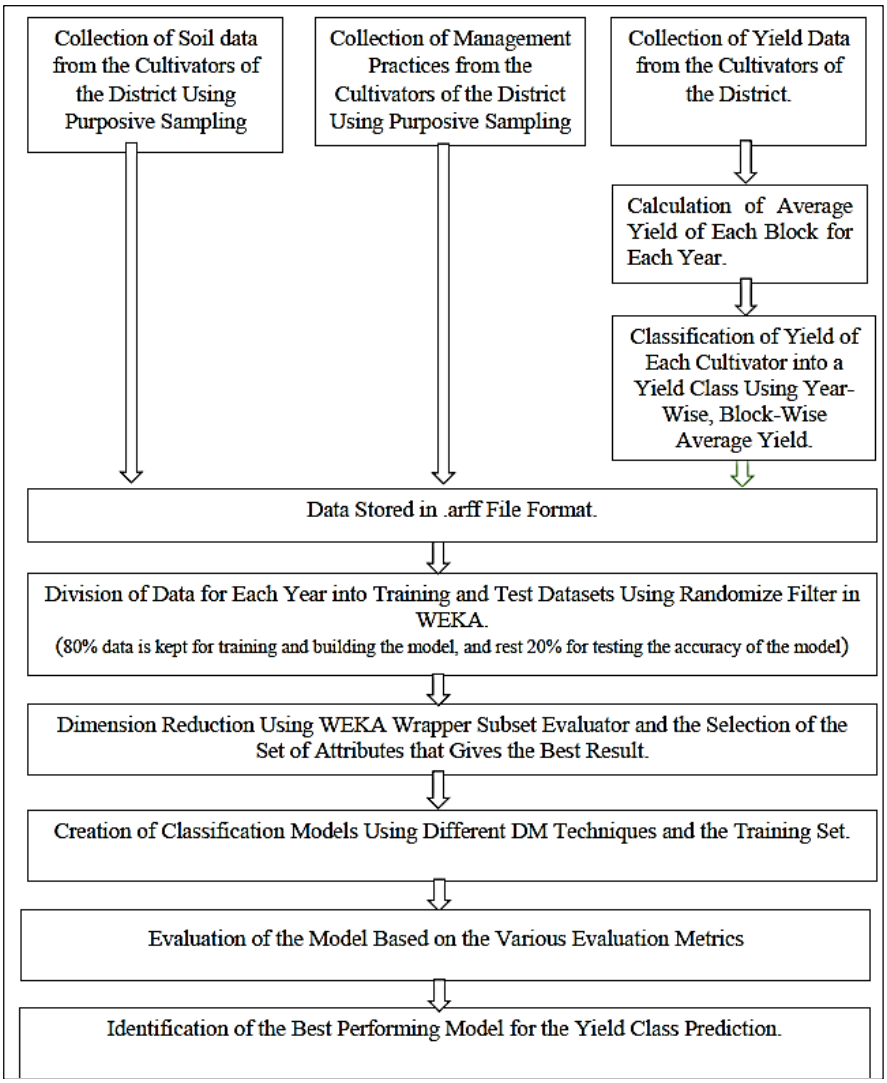

Fig 1: Yield Class Prediction model (YC model)

2.1 Data Collection

The target area of the Research is the cultivators in the Patiala district of Punjab, India. The research work starts with the collection of data pertaining to wheat crop for the period of five years i.e. from 2015 to 2019. Sampling method is applied to carry out the research and Stratified Sampling is used for the data collection. The strata are formed on the basis of Community Development blocks (CD blocks) of the district. The Patiala district is divided into eight CD blocks, namely Bhunerhedi, Ghanour, Nabha, Patiala, Patran, Rajpura, Samana and Sanour. The total number of samples are divided equally into these eight blocks. Purposive sampling is used to collect data from those cultivators who can provide the richest information and are willing to share information about the management practices followed by them, the soil type and the wheat yield. The data collection started in 2015 and continued through 2019 from the same set of cultivators. Besides this, the daily maximum temperature and the rainfall data for the Patiala weather station is acquired from the Indian Meteorological Department (IMD), Pune. The temperature, rainfall and yield data are used to create BAY model for the prediction of Block-wise average yield. The knowledge of the wheat phenology is used to divide the wheat growth period into four phases - the Germination phase, the Vegetative phase, the Reproductive phase, and the Grain Development and Ripening phase. The yearly average yield for a particular block is calculated by adding per acre yield of all the samples from that block and then by dividing the sum by the number of samples. The average temperature and the total rainfall during each phase are calculated and their effect on the yield is analysed. Stepwise Multiple Linear Regression is used to develop an Empirical model for the prediction of block-wise average wheat yield for the Patiala district [13].

The data pertaining to soil and the management practices along with the yield data is used to create the YC model. To find the variations in the yield of different cultivators, based on the soil in their farms and the management practices they follow, the variations caused by the other factors such as temperature and rainfall, and inter-block variation should be kept as constant. The yield of each cultivator is assigned to a yield-class based on the average yield (AY) of the block for that particular year. The Wheat dataset consists of a total of 1400 instances, which are divided into a training dataset and a test dataset. The Training dataset consists of 1120 instances and the Test dataset consists of 280 instances.

2.2 Selection of Factors

The questionnaire included the following factors related to soil and management practices:

Table 1: Agronomic Factors Affecting the Wheat Yield

S. No.	Factors Related to Soil	S. No.	Factors Related to Management Practices
1.	Soil Texture	1.	Number of Crops
2.	pH Level	2.	Crop Rotation
3.	Electrical Conductivity	3.	Seed Treatment
		4.	Sowing Date
		5.	Variety
		6.	Tillage
		7.	Farmyard Manure used?
		8.	Seed Rate
		9.	DAP
		10.	Urea
		11.	Irrigation Schedule
		12.	Weedicide used?
		13.	Fungicide used?
		14.	Insecticide used?

From the thorough study of the data, it is observed that some of the management practices are followed by almost all the cultivators in the region, and thus, such factors are not vital in the calculation of variance in the yield. The factors from the selected factors for the study, that fall under this category are number of crops, crop rotation, seed treatment, irrigation schedule, and the use of weedicide, Fungicide, and Insecticide. However, sowing date, variety, tillage, the use of farmyard manure, seed rate, DAP, and Urea are the important factors that apparently influence the yield.

It is also observed that on the one hand the cultivators were able to provide the information about the soil texture, but on the other hand, they lack the awareness about the pH level and Electrical Conductivity of the soil; so, the researcher could not collect the data of these two factors.

On the basis of the above observations, the following factors have been selected to calculate the variance of yield among the targeted cultivators:

Table 2: The Selected Agronomic Factors Affecting the Wheat Yield

S. No.	Factors Related to Soil	S. No.	Factors Related to Management Practices
1.	Soil Texture	1.	Sowing Date
		2.	Variety
		3.	Tillage
		4.	Farmyard Manure used?
		5.	Seed Rate
		6.	DAP
		7.	Urea

2.3 Data Pre-processing

To find the variations in the yield of different cultivators based on the soil in their farms and the management practices they follow, the variations caused by other factors such as temperature and rainfall, and inter-block variation should be kept as constant.

Table 3: Formulation of Yield-Class on the Basis of Average Yield

S. No.	Yield-Class	Criteria
1.	EL	Yield <= (AY - 4.5)
2.	L2	Yield >= (AY - 4.5) and Yield <= (AY - 3.5)
3.	L1	Yield >= (AY - 3.5) and Yield <= (AY - 2.5)
4.	L	Yield >= (AY - 2.5) and Yield <= (AY - 1.5)
5.	LM	Yield >= (AY - 1.5) and Yield <= (AY - 0.5)
6.	M	Yield >= (AY - 0.5) and Yield <= (AY + 0.5)
7.	HM	Yield >= (AY + 0.5) and Yield <= (AY + 1.5)
8.	H	Yield >= (AY + 1.5) and Yield <= (AY + 2.5)
9.	H1	Yield >= (AY + 2.5) and Yield <= (AY + 3.5)
10.	H2	Yield >= (AY + 3.5) and Yield <= (AY + 4.5)
11.	EH	Yield >= (AY + 4.5)

The yield of each cultivator is assigned to a yield-class based on the average yield (AY) of the block for that particular year. If the yield of the cultivator is equal to the average yield (AY) or in the range of 0.5 quintal above or below the AY, then it is assigned to the yield class - Moderate (M). Besides this, ten other classes are defined - five for yield above the AY and five for yield below the AY. The Yield Classes - EL through EH are defined as shown in Table 3.

```
@relation Wheat_Data

@attribute Block string
@attribute Year {2015,2016,2017,2018,2019}
@attribute SoilTexture {Maira,Daakar,Retli,Cheekni,Kallar}
@attribute SowingDate {D1,D2,D3}
@attribute Variety {HD-1105,HD-2733,HD-2967,HD-3086,Barbat}
@attribute Tillage {CTR,CTB,HS,SS}
@attribute FM {Yes,No}
@attribute SeedRate {35-38,39-42,43-46,47-50,>50}
@attribute DAP {31-40,41-50,51-60}
@attribute UREA {81-90,91-100,101-110,111-120,121-130,131-140}
@attribute YC {EL,L2,L1,L,LM,M,HM,H,H1,H2,EH}

@data
Nabha,2016,Retli,D1,HD-2967,HS,No,>50,31-40,101-110,18.7,LM
Nabha,2018,Cheekni,D1,HD-2967,SS,No,43-46,31-40,91-100,LM
Ghanour,2015,Maira,D1,HD-2967,CTR,No,43-46,41-50,91-100,H1
Rajpura,2019,Daakar,D2,HD-2733,CTR,Yes,43-46,31-40,91-100,HM
Bhunerheri,2019,Cheekni,D1,HD-3086,CTB,No,43-46,41-50,111-120,M
Patran,2019,Kallar,D2,HD-3086,CTR,No,43-46,41-50,121-130,L1
Sanour,2015,Kallar,D1,HD-2733,CTR,No,43-46,41-50,111-120,L
Bhunerheri,2017,Cheekni,D1,HD-2967,CTB,No,>50,41-50,111-120,LM
Sanour,2016,Daakar,D1,HD-2733,CTR,No,39-42,31-40,91-100,HM
Ghanour,2015,Daakar,D1,HD-1105,HS,No,>50,31-40,101-110,M
Patiala,2019,Cheekni,D1,HD-3086,CTR,Yes,43-46,31-40,91-100,H
Patran,2019,Retli,D2,Barbat,CTR,No,43-46,41-50,121-130,LM
Ghanour,2015,Cheekni,D1,HD-1105,SS,No,43-46,31-40,91-100,LM
Bhunerheri,2017,Daakar,D1,HD-2967,HS,No,>50,31-40,101-110,M
```

Fig 2: Wheat Dataset in ARFF Format

The data mining tool WEKA is used to carry out the Data Mining process. WEKA does not accept files in normal XLS or XLSX format. The default file format of WEKA is ARFF - "Attribute-Relation File Format", although it accepts files in comma-separated value (CSV) format, C4.5 format, ARFF file is an ASCII text file which gives a list of instances that share a set of attributes [3]. Owing to this, the wheat yield dataset is converted into ARFF format. A screen shot of data, thus prepared is shown in Figure 2.

For Supervised Learning problems, the performance of a technique is measured in terms of the error rate. The model predicts the class of each instance - if it is correct, it is counted as a success; if not, it is an error. The error rate is the proportion of errors made over a whole set of instances, and it measures the overall performance of the model. The model should be evaluated on the future performance and not the past performance on old data. Therefore, it is always better to evaluate a model with a test set which is different from the training set. If the same training set is used as the test set, the results can be misleading. So, given that there are enough instances, the dataset can be divided into a training set and a test set. In the current research, the Wheat

dataset has 1400 instances - so the data is divided into training and test dataset. The Wheat dataset is loaded into WEKA. The Randomise filter **(**Filters->Unsupervised->Instance->Randomise) is used to shuffle the order of instances of the dataset. The top 80% of the instances are saved in a file named - Training. arff and the bottom 20% of the instances are saved in the file named - Test. arff. The data in these files serves as training and test data, respectively.

2.4 Dimension Reduction

Dimension Reduction (or Feature Selection) is the process of reduction of factors (or selecting the most relevant features) in the dataset [19]. All the attributes (features or variables) of a high-dimensional dataset may not be of importance to understand the underlying phenomena of interest. So, it is desirable to reduce the number of independent variables and to remove those variables which do not substantially affect the dependent variable [20]. The use of dimensionality reduction and attribute selection techniques is informed by prior research that emphasizes eliminating irrelevant features to boost model interpretability and accuracy [21].

In this research, Wrapper method of attribute selection is used for dimension reduction. The Wrapper method is based on a specific machine learning algorithm. The Wrapper method follows a greedy search approach by evaluating all the possible combinations of attributes against the evaluation criterion. Here, the evaluation criterion is the percentage of instances that have been correctly classified by the model built by the learning algorithm. The best set of attributes may be found by using - Forward Search, Backward search or Bi-directional search mechanism.

Starting with an empty subset, the forward search evaluates all the possible single-attribute expansions to the current subset. The attribute which helps in achieving the best score is included permanently. When there is no further improvement in the accuracy of the classifier on the inclusion of an attribute from the remaining set of attributes, the search is terminated [21].

```
=== Run information ===
Evaluator:    weka.attributeSelection.WrapperSubsetEval -B weka.classifiers.trees.J48 -F 5 -T 0.01 -R 1 -E DEFAULT -- -C 0.25 -M 2
Search:       weka.attributeSelection.BestFirst -D 0 -N 5
Relation:     TrainingData
Instances:    1120
Attributes:   9
              SoilTexture
              SowingDate
              Variety
              Tillage
              FM
              SeedRate
              DAP
              UREA
              YC
Evaluation mode:    evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
        Best first.
        Start set: 1,2,3,4,5,6,7,8,
        Search direction: backward
        Stale search after 5 node expansions
        Total number of subsets evaluated: 51
        Merit of best subset found:    0.756

Attribute Subset Evaluator (supervised, Class (nominal): 9 YC):
        Wrapper Subset Evaluator
        Learning scheme: weka.classifiers.trees.J48
        Scheme options: -C 0.25 -M 2
        Subset evaluation: classification accuracy
        Number of folds for accuracy estimation: 5

Selected attributes: 1,2,3,4,5 : 5
                     SoilTexture
                     SowingDate
                     Variety
                     Tillage
                     FM
```

Fig 3: Attribute Selection

The Backward selection starts with a full set of attributes. It, then evaluates all the possible single-attribute depletions from the current subset. If the removal of an attribute improves the accuracy of the model, the attribute is permanently removed. This process is continued until the removal of an attribute starts decreasing the accuracy of the model. Bi-directional search mechanism starts in the same way as does forward selection, but while adding a new feature, it uses backward selection to check the merit of the already added attributes. If the exclusion of an attribute improves the accuracy of a newly formed set, then that attribute is removed.

In this study, J48 (Decision Tree) algorithm is used for attribute subset evaluation. The search starts with a full set of attributes; backward selection mechanism is used to select the set of attributes that gives the best result. The run-information of Wrapper method for attribute selection is shown in Figure 3. Thus, five attributes have been selected to predict the yield class. These attributes are soil texture, sowing date, variety, tillage and farmyard manure (FM) used.

2.5 The Selected Attributes

The selected attributes have been coded and given values to facilitate the further calculations and to develop the models. The following table gives the details of the codes given to the selected attributes:

Table 4: Domain of the Selected Attributes

Soil Texture	Sowing Date	Variety	Tillage	Farmyard Manure Used? (FM)
Maira	D1	HD-3086	CTB	Yes
Daakar	D2	HD-2967	CTR	No
Retli	D3	Barbat	SS	
Cheekni		HD-1105	HS	
Kallar		HD-2733		

The table clarifies that there are five types of Soil Texture in the district of Patiala i.e. Maira, Daakar, Retli, Cheekni, and Kallar; the sowing date is classified into three categories i.e. D1 (November 1 - November 15), D2 (November 16 - November 30), and D3 (After November 30); the most used variety of seeds in the district are HD-3086, HD-2967, Barbat, HD-1105, and HD-2733. The table also shows that the farmers used Conventional Tilling with Broadcast Planting (CTB), Conventional Tilling with Row Planting (CTR), Super Seeder (SS), and Happy Seeder (HS) as the methods of tillage, and that the use of Farmyard Manure is classified into Yes/No category.

2.6 Metrics Used

The detail of the Metrics used for evaluation of the Classifiers is already discussed in a study -, so only a brief summary is given here. To understand the different metrics used to evaluate the performance of a classifier, the primary requirement is to understand the confusion matrix. The size of the matrix depends on the number of classes in the dataset. For n classes, an n*n matrix will be created. The rows depict the actual class to which the instances belong, and columns depict the class predicted by the classifier. The number of correctly classified instances is given by the diagonal elements.

Table 5: Confusion Matrix

		Predicted class		
		A	**B**	**Actual Total**
Actual class	**A**	TP_A^*	FP_B	$TP_A + FP_B$
	B	FP_A	TP_B^*	$FP_A + TP_B$
	Predicted Total	$TP_A + FP_A$	$FP_B + TP_B$	

* Correctly classified Instances

In the above table, TP_A is the number of instances in the dataset that actually belong to class A and the classifier has also predicted their class as A. These are known as true positives for class A. FP_A is the number of instances in the dataset that have been predicted as belonging to class A but which actually belong to class B. These are known as false positives for class A. TP_B is the number of instances in the dataset that actually belong to class B and the classifier has also predicted their class as B. These are known as true positives for class B. FP_B is the number of instances in the dataset that have been predicted as belonging to class B but which actually belong to class A. These are known as false positives for class B. The metrics used for evaluation of classifiers are categorised into three types - Threshold Evaluation Metrics (TEM), Numerical Evaluation Metrics (NEM) and Build Time and Size Metrics (BTSM).

Table 6: Classification Metrics Categorisation

Threshold Evaluation Metrics (TEM)	Numerical Evaluation Metrics (NEM)	Build Time and Size Metrics (BTSM)
Percent Correct	Mean Absolute Error	Elapsed Time Training
True Positive Rate	Root Mean Squared Error	Serialised Model Size
False Positive Rate	Relative Absolute Error	
Precision	Root Relative Squared Error	
Recall		
F measure		
Kappa Statistic		
Area_ under_ ROC		

3. Results and Discussion

3.1 YC Model

The Training dataset created in Section 2.3 is used to create the models using different Data Mining techniques. Seven techniques - Naive Bayes, Decision Table (Rule Set Induction), IBk (Nearest Neighbor), SVM, J48 (Decision Tree), Random Forest, and Multilayer Perceptron (Neural Network) are used to create models using the Training dataset. The performance of the Models is evaluated using the test dataset, based on the above defined metrics.

WEKA Experimenter interface is used to carry out the evaluation of the different classification techniques used for building the models. Threshold Evaluation Metrics (TEM) for different techniques are given in Table 7.

For all the metrics, SVM is taken as the base technique. An x with the value of a metric indicates that the value of the metric is significantly below the value of the base technique. A v with the value of a metric indicates that the value is significantly more than the value of the base technique. For the metrics such as True Positive Rate which are evaluated for each class, weighted average of all the classes in the dataset is displayed. From the above table, it is clear that for this particular dataset, Naïve Bayes is the worst performer. Multilayer Perceptron has the best performance for most of the metrics but the performance of IBk (Nearest Neighbor), J48 (Decision Tree) and Random Forest is also comparable. WEKA is not giving any value for Weighted Average Precision and Weighted Average F measure for Naïve Bayes. Actually, Naïve Bayes does not classify any of the instances in “EH” class. So, value of (True Positives + False Positives) for this class comes out to be 0.

Table 4: Threshold Evaluation Metrics for Different Techniques

Metric	Naïve Bayes	Decision Table (Rule Set)	IBk (Nearest Neighbor)	SVM	J48 (Decision Tree)	Random Forest	Multilayer Perceptron (Neural Network)
Percent Correct	$27.14_{x}^{\#}$	67.14_{x}	78.57_{v}	72.14	76.79_{v}	78.57_{v}	81.79_{v}^{*}
Kappa Statistic	$0.15_{x}^{\#}$	0.63_{x}	0.75_{v}	0.68	0.73_{v}	0.75_{v}	0.79_{v}^{*}
Weighted Average True Positive Rate	$0.27_{x}^{\#}$	0.67_{x}	0.79_{v}	0.72	0.77_{v}	0.79_{v}	0.82_{v}^{*}
Weighted Average False Positive Rate	$0.12_{v}^{\#}$	0.03_{x}	0.03_{x}	0.04	0.03_{x}	0.03_{x}	0.02_{x}^{*}
Weighted Average Precision	$?^{\#}$	0.80_{v}	0.79_{v}	0.72	0.77_{v}	0.79_{v}	0.84_{v}^{*}
Weighted Average Recall	$0.27_{x}^{\#}$	0.67_{x}	0.79_{v}	0.72	0.77_{v}	0.79_{v}	0.82_{v}^{*}
Weighted Average F measure	$?^{\#}$	0.70_{x}	0.78_{v}	0.71	0.77_{v}	0.79_{v}	0.82_{v}^{*}
Weighted Average Area_under_ROC	$0.82_{x}^{\#}$	0.87_{x}	0.94	0.94	0.95_{v}	0.96_{v}^{*}	0.94_{x}

* Best Performing Technique for a Particular Metric
\# Worst Performing Technique for a Particular Metric
x Value for the metric significantly less than the Base Technique
v Value for the metric significantly more than the Base Technique

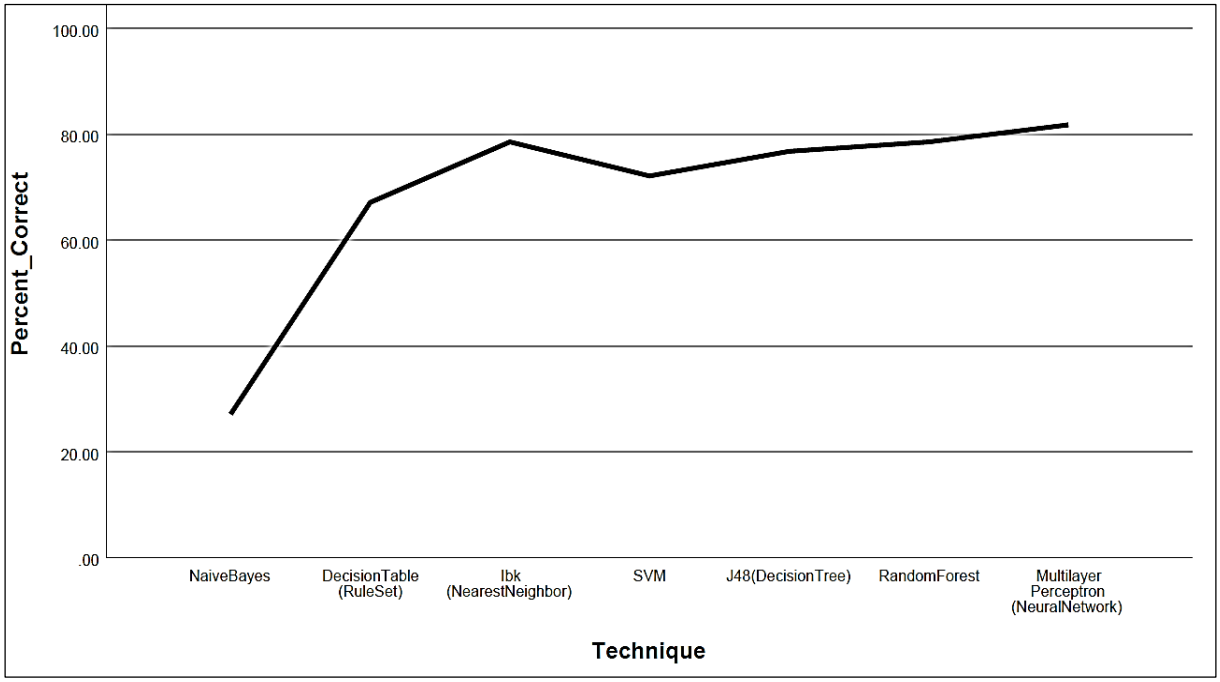

Fig 4: Percent of Correctly Classified Instances for Different Techniques

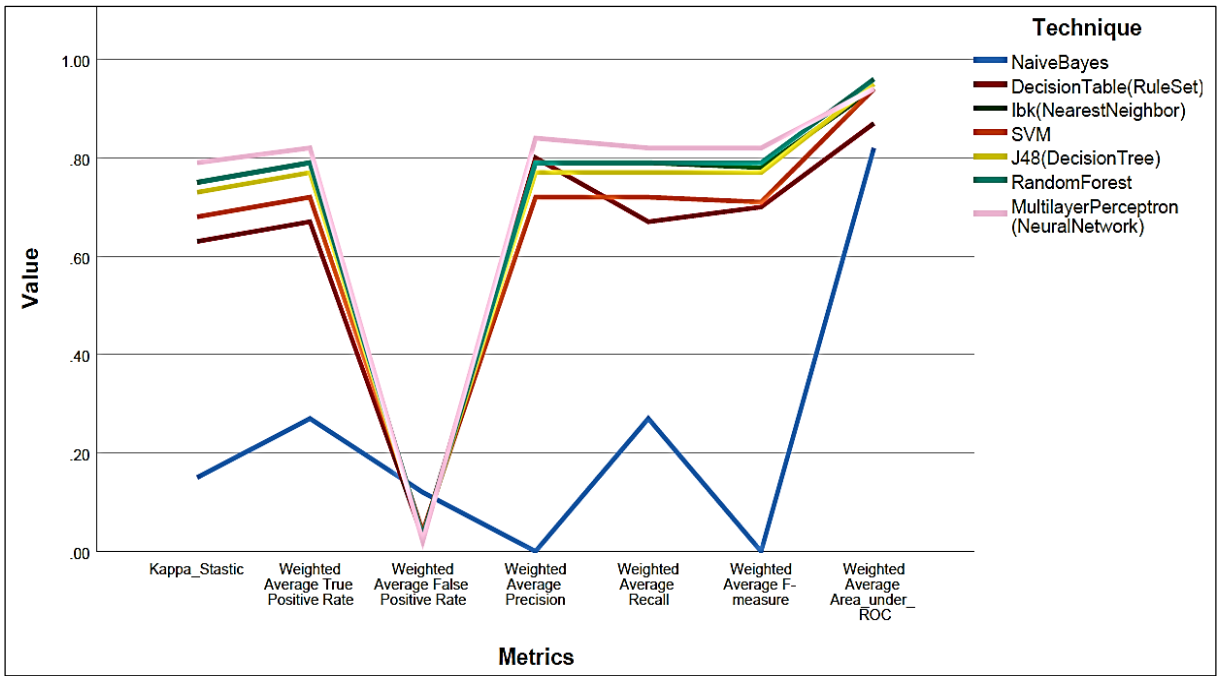

Fig 5: Threshold Evaluation Metrics for Different Techniques

The Percent Accuracy for different techniques is represented graphically in Figure 4 and the other Threshold Evaluation Metrics are shown graphically in Figure 5.

Numerical Evaluation Metrics are based on probabilistic understanding of error. They measure the deviation from the true probability . The error rate is evaluated by comparing the predicted and the actual values of instances in the test set. The results of Numeric Evaluation Metrics for different techniques are given below:

Table 8: Numeric Evaluation Metrics for Different Techniques

Metric	Naïve Bayes	Decision Table (Rule Set)	IBk (Nearest Neighbor)	SVM	J48 (Decision Tree)	Random Forest	Multilayer Perceptron (Neural Network)
Mean Absolute Error	0.14_x	0.14_x	0.04_x^*	$0.15^\#$	0.05_x	0.05_x	0.04_x^*
Root Mean Squared Error	0.26_x	0.25_x	0.17_x^*	$0.27^\#$	0.18_x	0.17_x^*	0.17_x^*
Relative Absolute Error	88.26_x	90.02_x	28.05_x	$95.25^\#$	31.16_x	31.50_x	25.92_x^*
Root Relative Squared Error	93.19_x	89.33_x	60.96_x	$94.59^\#$	63.37_x	59.43_x	58.78_x^*

* Best Performing Technique for a Particular Metric
\# Worst Performing Technique for a Particular Metric
x Value for the metric significantly less than the Base Technique
v Value for the metric significantly more than the Base Technique

The above table shows slightly surprising results when compared with the previous table. The Support Vector Machine classifier performed reasonably well when only threshold values are considered, but in this case, it is clearly the worst performer. It is closely followed by Decision Table and Naïve Bayes classifiers.

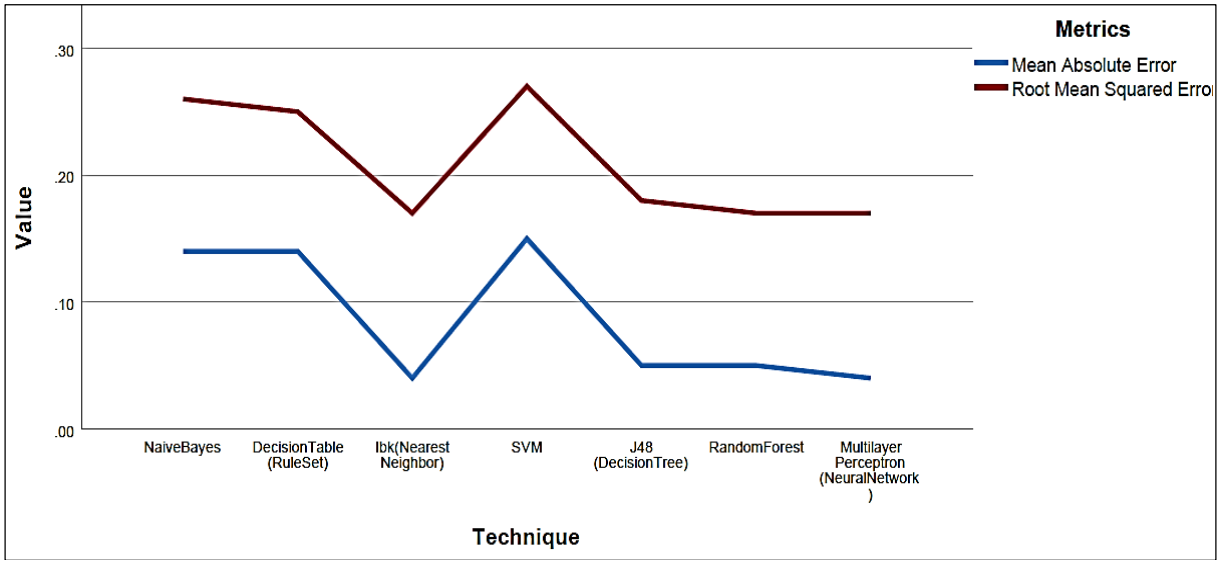

Fig 6: Mean Absolute Error and Root Mean Squared Error for Different Techniques

The Multilayer Perceptron is apparently the best performer as the predicted values show minimum deviation from the actual values. The performance of IBk, J48 and Random Forest classifiers is also comparable to that of Multilayer Perceptron. The graph for MAE and RMSE is shown in Figure 6, and the graph for RAE and RRSE is shown in Figure 7.

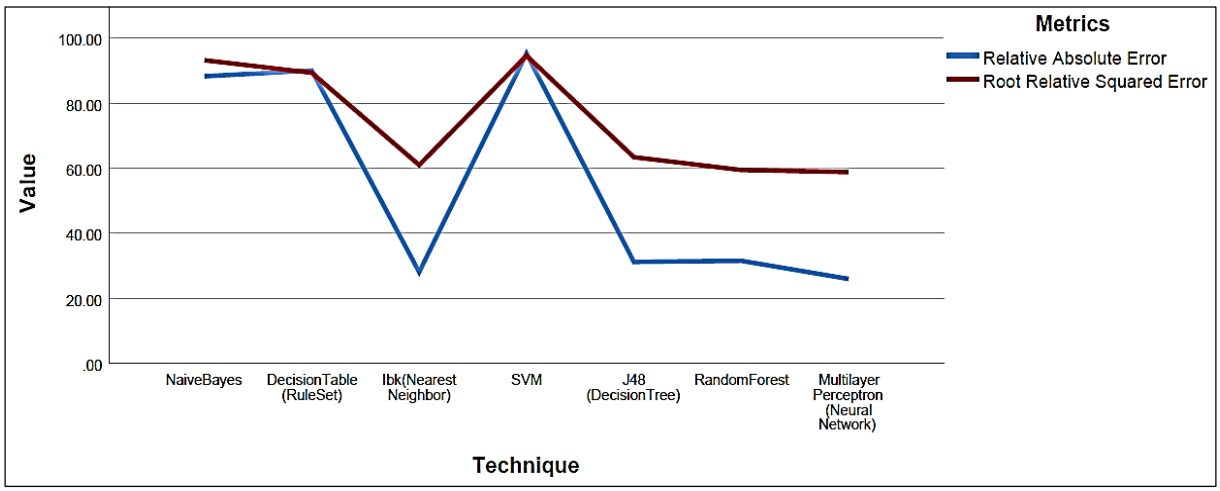

Fig 7: Relative Absolute Error and Root Relative Squared Error for Different Techniques

Build Time and Size Metrics give measure of the time required to build a model by a particular classifier and the size of the model which is developed. The Table below shows the time taken to build the model and size of the model developed for different techniques.

Table 9: Build Time and Size Metrics for Different Techniques

Metric	Naive Bayes	Decision Table (Rule Set)	IBk (Nearest Neighbor)	SVM	J48 (Decision Tree)	Random Forest	Multilayer Perceptron (Neural Network)
Elapsed Time Training (in seconds)	0.00_x	0.03_x	0.00_x^*	0.41	0.00_x^*	0.09_x	$3.23_v^\#$
Serialised Model Size (in Bytes)	7142_x^*	74676_v	90354_v	30241	125017_v	$5665122_v^\#$	40280_v

* Best Performing Technique for a Particular Metric
\# Worst Performing Technique for a Particular Metric
x Value for the metric significantly less than the Base Technique
v Value for the metric significantly more than the Base Technique

Hence, it is clear that Multilayer Perceptron takes highest time to build the model. SVM and Random Forest also take significantly more time than the rest of the techniques, though the time taken is much less than Multilayer Perceptron.

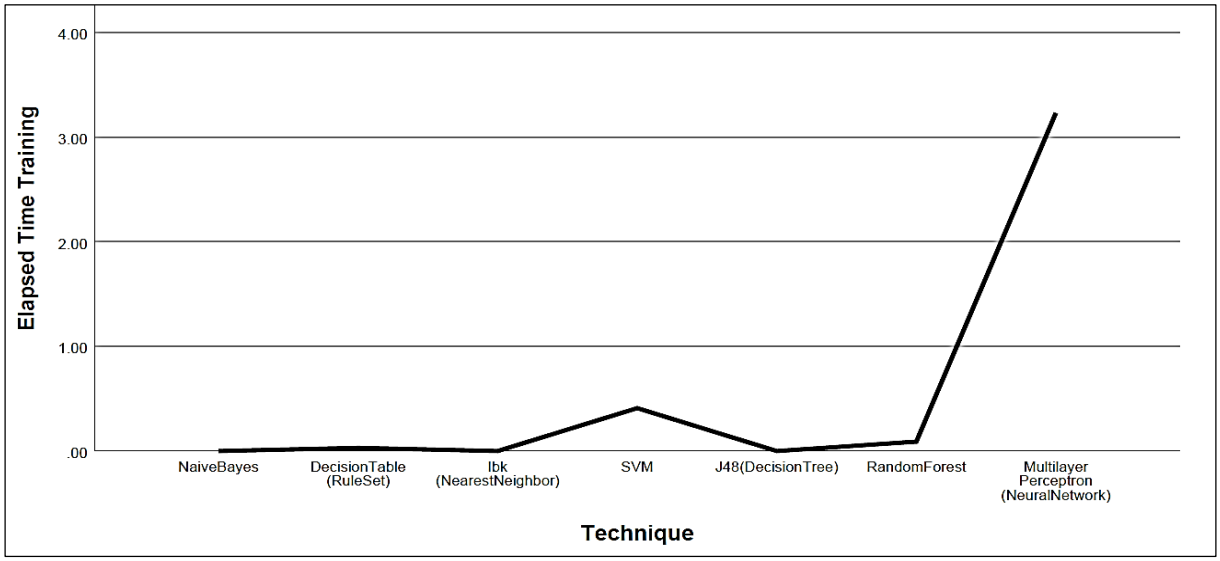

Fig 8: Elapsed Time Training for Different Techniques

Random Forest takes the largest amount of space to build the model. Random Forest Classifier is an ensemble learning technique which operates through the construction of multiple decision trees at the time of training. Random forest algorithm constructs various decision trees using random subset of data samples and then obtains the prediction from every tree. Finally, it chooses the best required solution with the help of voting. As Random Forest creates multiple models to make a prediction, it ends up using the largest amount of space. The memory requirement for other models is negligible as compared to Random Forest. Compared to rest of the techniques, the space taken by J48 and IBk is also significantly high. The smallest model size is that of Naïve Bayes followed by SVM and Multilayer Perceptron.

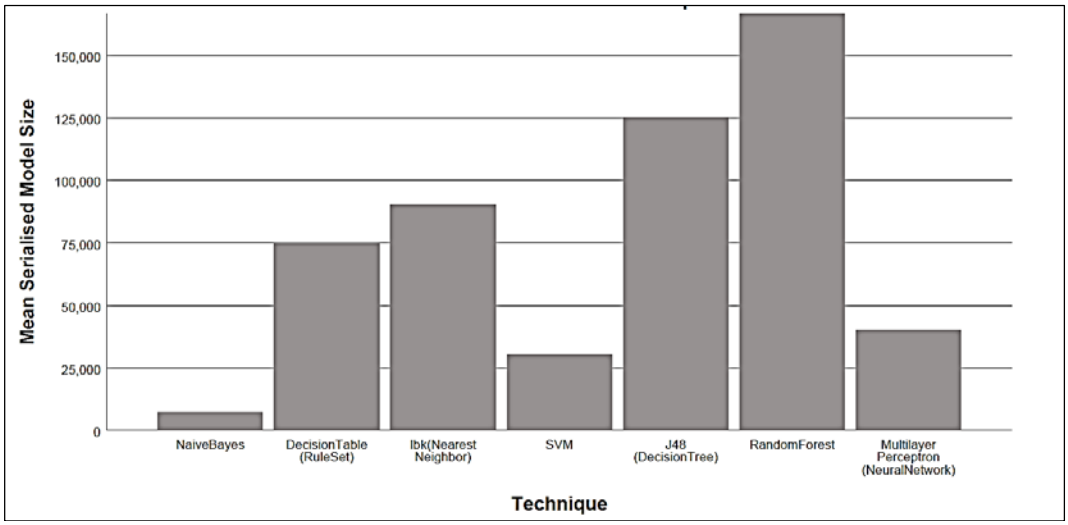

Fig 9: Serialised Model Size for Different Techniques

The performance of different techniques for different categories of metrics for Wheat Yield dataset is summarised in the graph shown in Figure 10.

This is apparent from the analysis of the graph given in Figure 10 that the Neural Networks technique gives the best performance for both TEM and NEM but it takes significantly more time than the other techniques to build the model. The Nearest Neighbor, Decision Tree Induction and Random Forest techniques also give good performance for TEM; for NEM, they closely follow Neural Networks and the time requirements are considerably less than the Neural Network technique. The space requirements of Random Forest are much greater than any other technique. Naïve Bayes gives the worst performance for TEM and also performs poorly for NEM, though its time and space requirements are minimal.

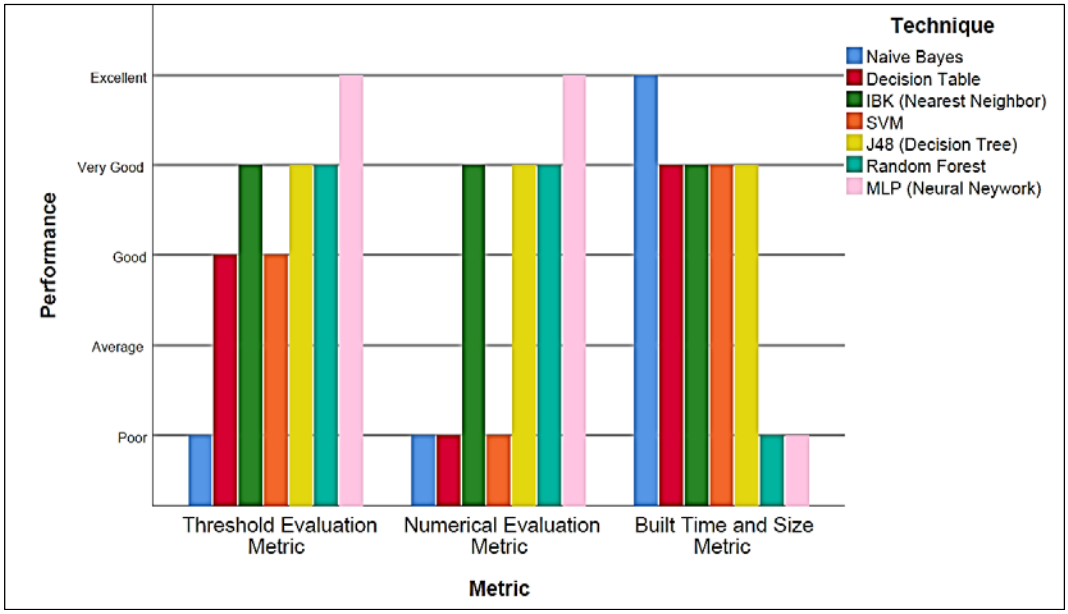

Fig 10: Performance of Different Techniques for Defined Metrics

The performance of Decision Table and SVM is much better than Naïve Bayes for TEM but SVM gives the worst performance for NEM and is closely followed by Naïve Bayes and Decision Table techniques. Only those techniques are studied further which give a good performance in TEM and NEM. So, we can leave Naïve Bayes, Decision Table and SVM techniques on the basis of their poor performance in TEM and NEM.

In case of misclassified instances, it also matters as to which class an instance has been assigned. The farther the class is from the actual class, greater is the error. We can categorise the misclassified instances as - the instances classified to the adjacent classes (a class one below or one above the actual class) and the instances classified to far-off classes (a class which is more than one class below or one class above the actual class). So, to further decide as to which technique to be adopted for prediction of Yield Class for the wheat dataset, the confusion matrices of the top four performing techniques are analysed. The rows of the confusion matrix depict the actual class of the instance and the columns depict the class predicted by a classifier. The instances that have been incorrectly classified to far-off classes are shown in red. The diagonal elements give the number of correctly classified instances. The confusion matrices for the four techniques - IBk (Nearest Neighbor), J48 (Decision Tree), Random Forest and MLP (Neural Network) are shown below.

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
4	1	0	2	0	0	0	0	0	0	0	a = EL
4	3	1	0	0	2	2	0	0	0	0	b = L2
0	0	17	2	2	0	0	1	0	0	0	c = L1
0	0	0	26	6	2	1	0	0	1	0	d = L
0	0	0	0	37	5	0	1	0	0	0	e = LM
0	0	0	1	0	44	2	2	1	0	0	f = M
0	0	0	0	1	0	42	4	1	1	0	g = HM
0	0	0	0	0	0	1	27	3	1	1	h = H
0	0	0	0	0	0	1	4	13	1	0	i = H1
0	0	0	0	0	0	0	0	1	5	0	j = H2
0	0	0	0	0	0	0	0	0	1	2	k = EH

Fig 11: Confusion Matrix for IBk (Nearest Neighbor)

The IBk (Nearest Neighbor) classified 220 instances correctly and 60 incorrectly. Out of the 60 incorrectly classified instances, 37 were assigned to the adjacent classes and 23 were assigned to far-off classes.

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
5	1	0	1	0	0	0	0	0	0	0	a = EL
5	4	2	1	0	0	0	0	0	0	0	b = L2
0	2	17	2	1	0	0	0	0	0	0	c = L1
0	0	1	26	7	1	1	0	0	0	0	d = L
0	0	0	1	34	7	0	1	0	0	0	e = LM
0	0	0	0	0	43	4	2	1	0	0	f = M
0	0	0	0	0	2	41	4	1	1	0	g = HM
0	0	0	0	0	0	4	24	3	1	1	h = H
0	0	0	0	0	0	1	4	14	0	0	i = H1
0	0	0	0	0	0	0	0	1	5	0	j = H2
0	0	0	0	0	0	0	0	0	1	2	k = EH

Fig 12: Confusion Matrix for J48 (Decision Tree)

The J48 (Decision Tree) technique classified 215 instances correctly and 65 incorrectly. Out of the 65 incorrectly classified instances, 52 were assigned to the adjacent classes and 13 were assigned to far-off classes.

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
5	1	0	1	0	0	0	0	0	0	0	a = EL
5	4	2	1	0	0	0	0	0	0	0	b = L2
0	0	17	3	1	1	0	0	0	0	0	c = L1
0	2	1	27	4	1	1	0	0	0	0	d = L
0	0	1	0	37	4	0	1	0	0	0	e = LM
0	0	0	0	0	44	3	2	1	0	0	f = M
0	0	0	0	1	0	39	7	1	1	0	g = HM
0	0	0	0	0	0	1	27	3	1	1	h = H
0	0	0	0	0	0	1	3	13	2	0	i = H1
0	0	0	0	0	0	0	0	1	5	0	j = H2
0	0	0	0	0	0	0	0	0	1	2	k = EH

Fig 13: Confusion Matrix for Random Forest

The Random Forest technique classified 220 instances correctly and 60 incorrectly. Out of the 60 incorrectly classified instances, 41 were assigned to the adjacent classes and 19 were assigned to far-off classes.

a	b	c	d	e	f	g	h	i	j	k	<-- classified as
5	1	0	0	0	1	0	0	0	0	0	a = EL
1	10	0	0	1	0	0	0	0	0	0	b = L2
0	4	15	1	0	2	0	0	0	0	0	c = L1
0	0	0	28	3	3	1	0	1	0	0	d = L
0	0	0	0	36	5	1	0	1	0	0	e = LM
0	0	0	0	0	44	3	2	1	0	0	f = M
0	0	0	0	0	0	40	7	1	1	0	g = HM
0	0	0	0	0	0	0	30	1	1	1	h = H
0	0	0	0	0	0	0	5	14	0	0	i = H1
0	0	0	0	0	0	0	0	1	5	0	j = H2
0	0	0	0	0	0	0	0	0	1	2	k = EH

Fig 14: Confusion Matrix for MLP (Neural Network)

The MLP (Neural Network) technique classified 229 instances correctly and 51 incorrectly. Out of the 51 incorrectly classified instances, 33 were assigned to the adjacent classes and 18 were assigned to far-off classes.

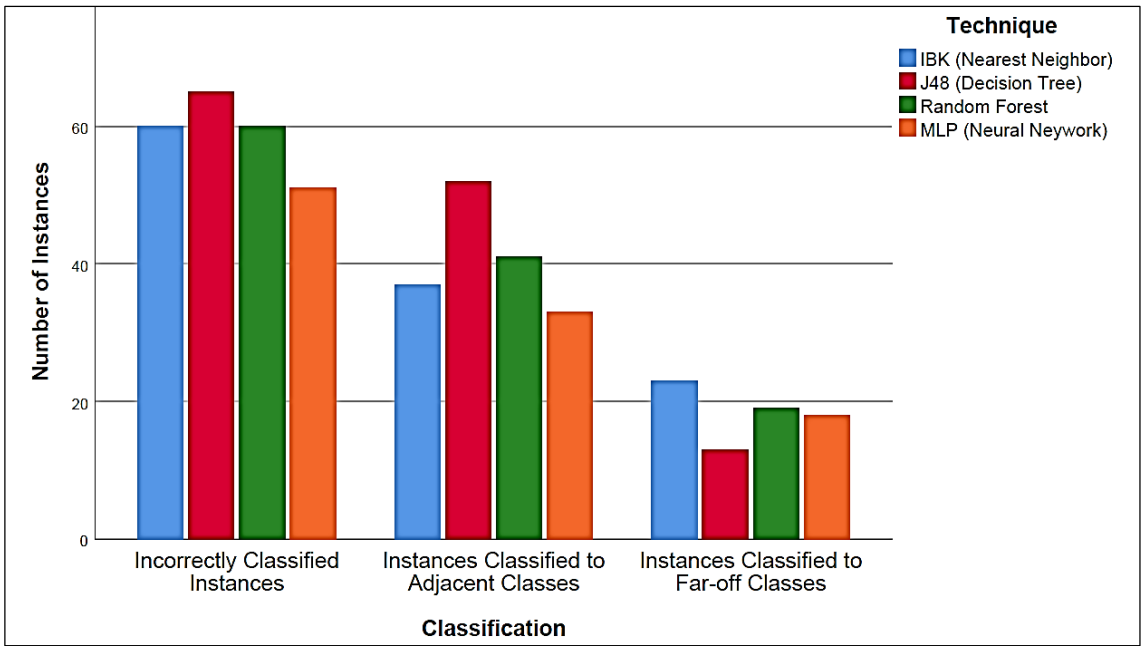

Fig 15: Classification of Incorrectly Classified Instances

The graph above shows the total number of incorrectly classified instances, the number of instances that have been incorrectly classified to the adjacent classes and the number of instances that have been incorrectly classified to far-off classes. It is clear from the Graph that the J48 techniques has the maximum number of incorrectly classified instances and MLP has minimum number of incorrectly classified instances, while IBk and Random Forest lie in-between these two with 60 incorrectly classified instances each. But for J48, most of the incorrectly classified instances have been assigned to the adjacent classes. IBk leads in having the greatest number of instances being assigned to far-off classes while Random Forest and MLP closely follow IBk with 19 and 18 instances incorrectly classified to far-off classes, respectively. Here, MLP is the most efficient technique in terms of maximum number of correct classifications, but it has a large number of instances which have been classified to far-off classes; whereas J48 has minimum efficiency in terms of correct classifications but most of the incorrectly classified instances have been assigned to the adjacent classes. So, it is not possible to choose any one technique at this stage and all the four techniques are used in the next stage of final yield estimation, so as to check that for which technique the percent error is minimum.

3.2 Development of Final Yield Prediction Model

The current section, explains the creation of Final Yield Prediction Model (FYP Model) through the integration of the Block-wise Average Prediction Model (BAY model) and the Yield Class Model (YC Model). As discussed in previous section, the BAY model is developed using the Data Mining tool - SPSS; the model is created based on the meteorological factors, and predicts the Block-wise Average Yield. Whereas the YC Model, developed in Section 3, is developed using Data Mining tool - WEKA; this model predicts the Yield Class of a particular cultivator based on the agronomic factors such as soil texture and management practices. The final stage in the study is the integration of these two models to predict the actual wheat yield of a cultivator. The Final Yield Prediction model (FYP model) takes Block-wise Average Yield (AY) predicted by the BAY model and Yield Class (YC) predicted by the YC model as input and makes the final yield prediction for a particular farmer.

The process followed by FYP model to predict the final yield is shown in Figure 16. When the yield of a particular cultivator in Patiala district for a particular year is to be predicted, the data pertaining to temperature, rainfall and block-name is given to the BAY model as input. The BAY model predicts the Average Yield for that particular block. The data pertaining to soil and management practices followed is given to the YC model as input. The YC model predicts the Yield Class of that cultivator. The Average Yield (AY) and Yield Class (YC) are integrated by the model to predict the final yield of the cultivator in the following way:

The Moderate class (M) is taken as the standard class. The value assigned to the standard class is equivalent to the Average Yield (AY) as calculated by the BAY model. If the Yield Class is ‘LM’, which is immediately below the ‘M’ class, then the Predicted Yield is calculated by subtracting one from the Average Yield. For each subsequent class below, one quintal more is subtracted to calculate the Predicted Yield. If the Yield Class is ‘HM’, which is immediately above the ‘M’ class, then the Predicted Yield is calculated by adding one to the Average Yield. For each subsequent class above, one quintal more is added to calculate the Predicted Yield.

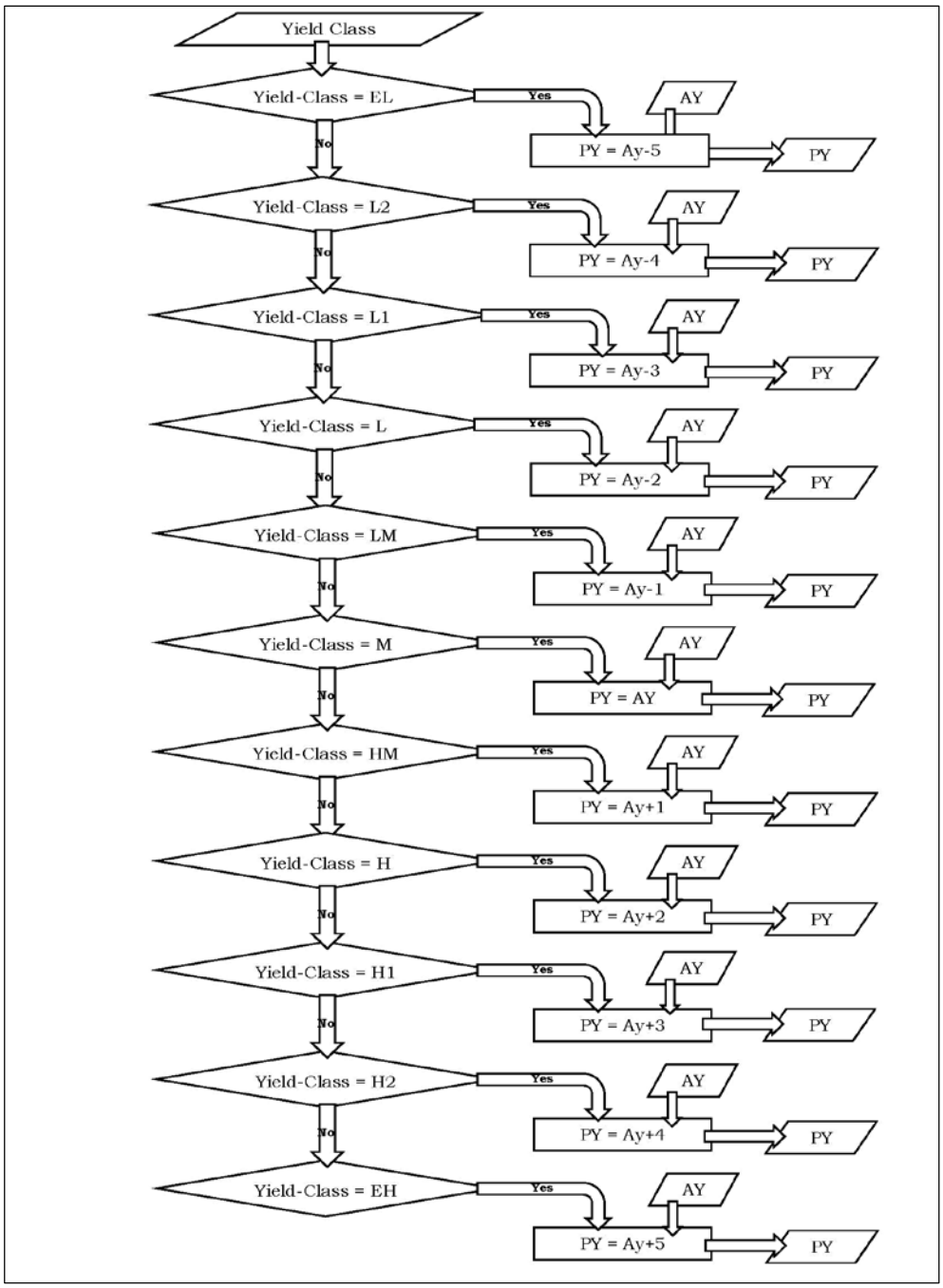

Figure 1: Final Yield Prediction Model (FYP model)

3.3 Performance Evaluation of FYP Model for the Selected Techniques

In the previous section, seven techniques have been compared so as to choose the best one for the creation of YC model. From these seven techniques, three techniques namely - Naïve Bayes, Decision Table and SVM are eliminated on the basis of their low performance in Threshold Evaluation Metric (TEM) and Numerical Evaluation Metric (NEM). But no conclusion could be reached so far as the remaining four techniques, namely - IBk (Nearest Neighbor), J48 (Decision Tree), Random Forest and MLP (Neural Network) are concerned. One technique has the best performance in TEM and NEM, while the others have lower number of instances misclassified to far-off classes. So, each of these four techniques are used to create YC model and to predict the Yield Class. The predicted yield class by each of these YC models is integrated with the Average Yield predicted by BAY model, to create FYP model which predicts the final yield. A comparative analysis of these predicted yield is made to finalise that which technique should be selected for the prediction of the final yield. Thus, the percent accuracy of these models is compared with each other. The graphic representation of the percent accuracy of these models is given below:

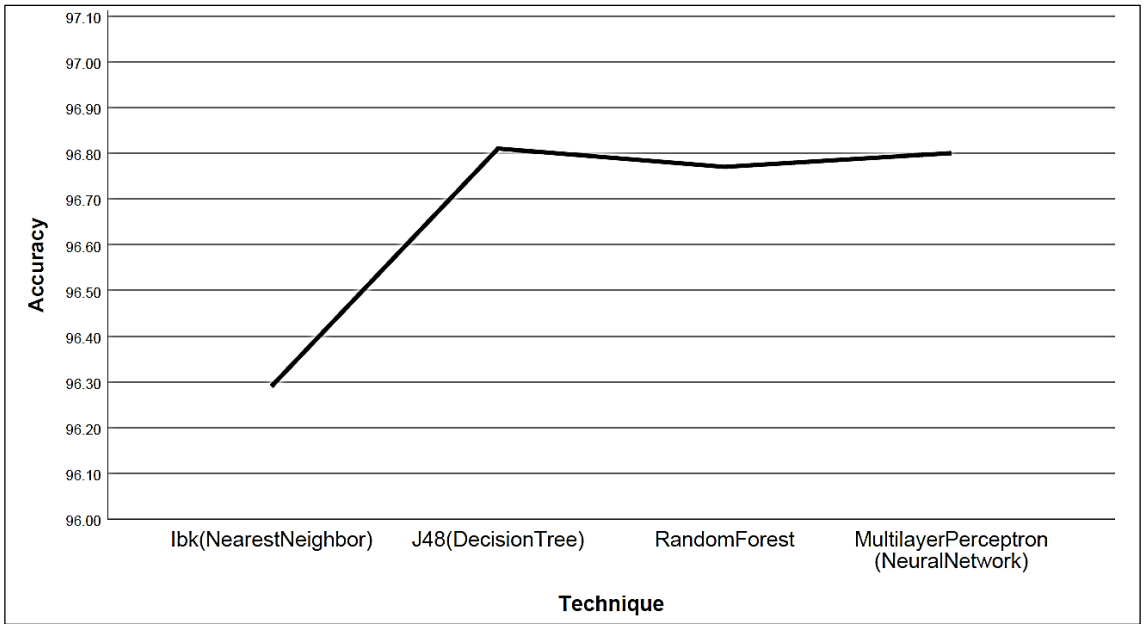

Fig 17: Final Yield Prediction Accuracy for the Selected Techniques

J48 (Decision Tree) leads the percent accuracy with 96.81 and it is closely followed by MLP (Neural Network) at 96.80 and Random Forest at 96.78 while IBk (Nearest Neighbor) has the lowest percent accuracy of 96.29. A further analysis is made to study the variation of percent error of predicted yield for these techniques. This variation is shown in the Figure below:

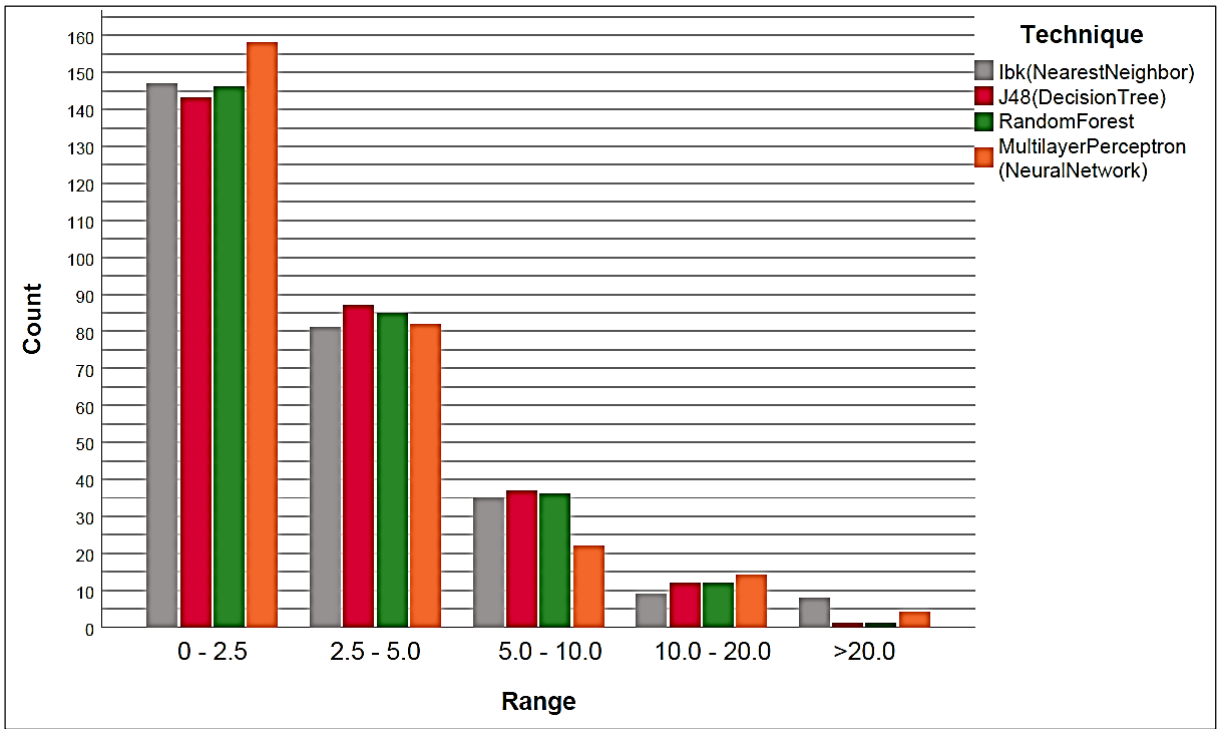

Fig 18: Variation in Percent-error of the Predicted Yield for the Selected Techniques

IBk (Nearest Neighbor) has maximum number of instances (eight) where the percent error is greater than 20. MLP (Neural Network) has four instances where the percent error is greater than 20, but has the greatest number of instances with the percent error greater than 10. Both J48 (Decision Tree) and Random Forest have maximum number of instances with the percent error less than 10 percent. Both have the least number of instances where the percent error is greater than 10 percent. This is the reason that J48 has the best percent accuracy in spite of getting the maximum number of incorrectly classified instances. It is more acceptable to have a larger number of instances with smaller error than to have a smaller number of instances with greater error. Though the Percent Accuracy of J48 (Decision Tree), Random Forest and MLP (Neural Network) is at par, yet J48 (Decision Tree) and Random Forest have much smaller number of instances that have the error of greater than 10 percent. So, these two techniques are chosen over MLP (Neural Network) to create the model. Amongst these two, the Random Forest is much more complex, as it uses ensemble learning technique which operates through the construction of multiple decision trees at the time of

training. The performance of Random Forest in Build Time and Model Size metric is much poorer than J48 (Decision Tree). So, J48 (Decision Tree) technique is chosen to create YC model for the prediction of the Yield Class.

3.4 Percent Accuracy of the FYP Model

The final Yield is predicted by integrating the average yield calculated on the basis of the selected meteorological variables and the block-name using BAY model, and Yield Class predicted by YC model using J48 (Decision Tree) technique on the basis of selected agronomic variables. The accuracy of the FYP model is tested on the test dataset previously created and the model is found to predict the wheat yield with an average accuracy of 96.81 percent. The graph of the number of instances along with the percent error is shown in the Figure below:

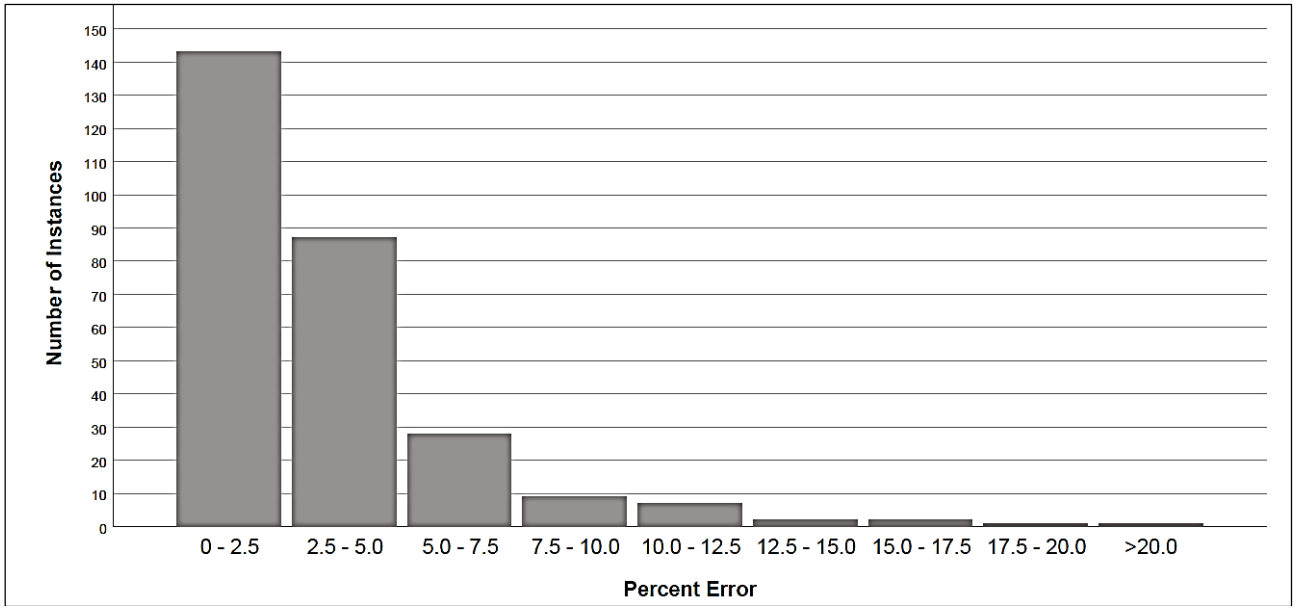

Fig 19: Percent Error of the Predicted Yield

Out of a total of 280 instances in the test dataset, the percent error of 143 instances is less than 2.5 percent, 267 instances have an error rate of lower than 10 percent, 13 instances have an error of more than 10 percent and there is only one instance with an error of more than 20 percent.

4. Conclusion

The study develops a model for wheat yield prediction based on the various meteorological and agronomic variables. The study found that the wheat yield is affected by several agronomic factors such as soil type and date of sowing, and meteorological factors such as temperature and rainfall. The study divides the set of factors into two categories - the factors which are responsible for year-wise variation in yield, and the factors which are responsible for the individual variation of yield for a particular year among various cultivators. Accordingly, two models - the Block-wise Average Yield Prediction model (BAY model) and the Yield Class Prediction model (YC model) are developed; the first model predicts the Block-wise Average Yield based on temperature, rainfall and the yield data, and the second predicts the Yield Class based on soil, management practices and yield data. Finally, these models, i.e., the BAY model and the YC model are integrated into the Final Yield Prediction model (FYP model) to predict the final yield of a particular cultivator.

The Wheat dataset consists of a total of 1400 instances, which are divided into a training dataset and a test dataset. The Training dataset consists of 1120 instances and the Test dataset consists of 280 instances. The performance of the proposed YC model is evaluated using different classifiers - Naïve Bayes, Decision Table, IBk (Nearest Neighbor), SVM, J48 (Decision Tree), Random Forest and MLP (Neural Network). For quantitative evaluation, three types of metrics - Threshold Evaluation Metrics (TEM), Numerical Evaluation Metrics (NEM) and Build Time and Size Metrics (BTSM) are defined. IBk (Nearest Neighbor), J48 (Decision Tree), Random Forest and MLP (Neural Network) give a good performance in terms of TEM and NEM. So, only these techniques are carried forward for the further study. The study shows that the J48 (Decision Tree) classifier, when used in the YC model, provides a balance between prediction accuracy and computational efficiency, outperforming other classifiers like IBk (Nearest Neighbor), Neural Networks and Random Forest in terms of build time, model size, and error rate. The FYP model achieves an impressive average accuracy of 96.81% when tested on a test dataset, confirming its reliability and effectiveness in predicting wheat yield.

The study holds significant practical use for farmers and policymakers. For farmers, the FYP model will be very much beneficial to predict an approximation of the yield they are going to achieve in the ongoing season. Based on the prediction, they may follow the management practices in a certain way that may help them to achieve a desired wheat yield. Yield prediction is also beneficial for the State in Forward marketing and in making policies for food security. Overall, this research contributes to the field of agricultural data mining by providing a comprehensive framework for yield prediction that can be adapted to other crops and regions. Future work could explore the integration of additional factors such as pest incidence, soil moisture levels, and remote sensing data to further refine the model and enhance its applicability in diverse agricultural contexts.

5. References

1. Chlingaryan A, Sukkarieh S, Whelan B. Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture: A review. Computers and Electronics in Agriculture. 2018;151:61-69.
2. Witten I, Frank E, Hall M, Pal C. Practical machine learning tools and techniques. Amsterdam, The Netherlands: Elsevier; 2005.
3. Bouckaert RR, Frank E, Hall M, Kirkby R, Reutemann P, Seewald A, *et al*. WEKA Manual for Version 3-8-1. Hamilton, New Zealand: University of Waikato; 2016.
4. Bassine F, Epule T, Kechchour A, Chehbouni A. Recent applications of machine learning, remote sensing, and IoT approaches in yield prediction: a critical review. arXiv preprint arXiv:2306.04566. 2023.
5. Attwal KPS. Design and development of a model for wheat yield prediction using data mining [dissertation]. Patiala: Punjabi University; 2020.
6. Attwal KPS, Dhiman AS. Investigation and comparative analysis of data mining techniques for the prediction of crop yield. Int J Sustain Agric Manag Inform. 2020;6(1):43-74.
7. Attwal KPS, Dhiman AS. A study of wheat morphology for yield prediction. Int J Res Agron. 2024;7(5):346-349.
8. Attwal KPS, Dhiman AS. A study of wheat phenology and the factors affecting the wheat yield. Int J Agric Extens Soc Dev. 2024;7(7):569-575.
9. Mishra S, Paygude P, Chaudhary S, Idate S. Use of data mining in crop yield prediction. In: 2nd International Conference on Inventive Systems and Control (ICISC); 2018.
10. Nath S, Debnath D, Sarkar P, Biswas A. Design of intelligent system in agriculture using data mining. In: International Conference on Computational Intelligence & IoT (ICCIIoT); Agartala; 2018.
11. Ghadge R, Kulkarni J, More P, Nene S, R PR. Prediction of crop yield using machine learning. Int Res J Eng Technol (IRJET). 2018;5(2):2237-2239.
12. Khaki S, Wang L. Crop yield prediction using deep neural networks. Front Plant Sci. 2019;10:621.
13. Attwal KPS, Dhiman AS. Mining effect of temperature and rainfall to develop an empirical model for wheat yield prediction. Recent Adv Comput Sci Commun. 2021;14(7):2195-2209.
14. Gupta S, Mohanty S, Behera DK. AI-based yield prediction: a thorough review. Indian J Sci Technol. 2025;18(10):822-838.
15. Akcapınar MC, Çakmak B. Yield prediction models for some wheat varieties with satellite-based drought indices and machine learning algorithms. Irrig Drain. 2025;74(1):237-250.
16. Li H, Gao J, Guo Y, Yuan XG. Application of XGBoost model and multi-source data for winter wheat yield prediction in Henan Province of China. Big Data Inf Anal. 2025;9:29-47.
17. Yang G, Jin N, Ai W, Zheng Z, He Y, He Y. Integrating remote sensing data assimilation, deep learning and large language model for interactive wheat breeding yield prediction. arXiv preprint arXiv:2501.04487. 2025.
18. Zhao Y, Du X, Xu J, Li Q, Zhang Y, Wang H, *et al*. Integrating WOFOST and deep learning for winter wheat yield estimation in the Huang-Huai-Hai Plain. Agriculture. 2025;15(12):1234-1245. *(Note: Page range assumed as missing)*
19. Yang P, Liu W, Zhou BB, Chawla S, Zomaya AY. Ensemble-based wrapper methods for feature selection and class imbalance learning. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining; Berlin, Heidelberg; 2013.
20. Fodor IK. A survey of dimension reduction techniques. CA (US); 2002.
21. Gutlein MFEHMKA. Large-scale attribute selection using wrappers. In: IEEE Symposium on Computational Intelligence and Data Mining; 2009.
22. Attwal KPS, Dhiman AS. Exploring data mining tool-Weka and using Weka to build and evaluate predictive models. Adv Appl Math Sci. 2020;19(6):451-469.
23. Ferri C, Hernandez-Orallo J, Modroiu R. An experimental comparison of performance measures for classification. Pattern Recognit Lett. 2009;30(1):27-38.
24. Venkaiah SA, Sunitha KS. A new approach for predicting crop yield prediction using data mining techniques. Int J Eng IT Sci Res (IJEISR). 2019;1(3):42-48.