SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V10I12P105

Volume 10 Issue 12, 45-61, December 2023

Original Article

A Hybrid Machine Learning Model with Combined Wrapper Feature Selection Techniques to Improve the Yield of Paddy

S. Muthukumaran[1], K. John Peter[2], E. Dilipkumar[3], S. Savithri[4], K. Senbagam[5]

[1,3,4]*Department of Master of Computer Applications, Dhanalakshmi Srinivasan College of Engineering and Technology, Tamilnadu, India.*
[2,5]*Department of Computer Science and Engineering, Dhanalakshmi Srinivasan College of Engineering and Technology, Tamilnadu, India.*

[1]*Corresponding Author : muthumphil11@gmail.com*

Received: 07 October 2023 Revised: 19 November 2023 Accepted: 09 December 2023 Published: 23 December 2023

Abstract - *A third of the earth's surface is taken up by agriculture, which is essential to the food production process. Paddy seeds are used to grow rice, which is a dependable food that is consumed by approximately half of all people worldwide. The alarming rate of population expansion makes it necessary for us to secure food security, and the nation should implement the measures required to increase the production of food grains. Since climatic, agronomic, irrigational, and cultivation techniques all affect paddy's growth. The goal of the study is to increase the production of rice by using Machine Learning (ML) techniques to forecast the variables that affect paddy growth. More attributes will be employed to build the dataset as ML techniques are used in real-time, which will reduce model performance, raise computing costs, and make the dataset more susceptible to overfitting. This research developed a Hybrid Machine Learning Model with Combined Wrapper Feature Selection Techniques (HMLCWFS) for forecasting paddy production to get over these challenges. The suggested approach selects the most significant features from the Paddy Dataset (PD) using five Feature Selection (FS) approaches: Backward Elimination (BE), Stepwise Forward Selection (SFS), Feature Importance (FI), Exhaustive FS (EFS), and Gradient Boosting (GB) approaches. Using Poincare's formula, the attributes chosen from each FS approach were concatenated, and the dataset was then recreated. The reconstructed dataset was used to deploy ML approaches like Decision Tree (DT) and Random Forest (RF), and the knowledge gleaned in the form of association rules was utilized to provide advice to paddy growers on how to increase productivity. The suggested model also takes into account the farmers' preferred paddy farming techniques and makes recommendations regarding which paddy variety they should cultivate. This is accomplished by connecting the input parameters to the real-time PD trained by employing Support Vector Machine (SVM), K-Nearest Neighbor (KNN), and Naive Bayes (NB) methods. The classifier's results were compared using performance metrics, and the findings demonstrate that the combined FS strategies employed in this research help to identify the elements contributing to the paddy crop's improvement.*

Keywords - *Feature Selection, Supervised Machine Learning, Paddy cultivation, SVM, Decision Tree, KNN.*

1. Introduction

Paddy is an essential cereal crop that is very important to the world's food systems and agriculture. Known by most as rice, this staple grain provides billions of people worldwide with its primary source of nutrition.

Paddy is widely known for its versatility and ability to flourish in a variety of growing circumstances, having been cultivated in a wide range of climates and geographies. Its complex farming methods are frequently fashioned by millennia of agricultural ingenuity and wisdom. A staple of many cultures and cuisines, rice is essential to tackling the world's problems with nutrition and food security. India's rice-growing regions range from lush plains to mountainous topography. Because of its broad dispersion, there is less chance of localized crop failures or unfavourable weather in any one area. Paddy is a crop that requires a lot of water and grows best in warm, humid climates [1].

It is primarily grown during the monsoon season of June through September, often known as the Kharif season. India produces a large variety of hybrid and traditional paddy kinds. Each state has its favourite cultivars that are best suited to the agro-climatic conditions there. Andhra Pradesh, Tamil Nadu, Uttar Pradesh, Telangana, Bihar, Odisha, and West Bengal are some of the states in India that produce the

most paddy. Although irrigation is essential to maintaining steady yields, rainfall is the main factor in paddy agriculture in India. Irrigation techniques using both surface and groundwater are employed.

The System of Rice Intensification (SRI), conventional flooded fields, and direct sowing techniques are some of the strategies that can be used to develop paddy. A large section of the rural population finds employment in the labour-intensive field of paddy production.

This covers agricultural employees who perform tasks including planting, weeding, harvesting, and post-harvest processing, in addition to farmers. Millions of people make their living from this industry, especially in rural areas. A considerable portion of India's Gross Domestic Product (GDP) comes from paddy, one of the nation's leading agricultural products. The addition of value to the total economic output comes from the cultivation, processing, and related operations of paddy [2].

Water is essential for paddy farming, and inadequate irrigation or water scarcity may result in a negative impact on yields. Both water waste and water contamination can result from ineffective water management techniques. The brown plant hopper and the blast disease are only two examples of the many pests and illnesses that can affect paddy crops. To control these dangers, it is frequently necessary to employ pesticides, which can be costly and harmful to the environment. India still uses human labour for many farming tasks, from planting to harvesting, in its paddy-growing regions. The lack of mechanization raises the need for labour and may result in expensive labour. Paddy agriculture may be impacted by erratic weather patterns, including variable rainfall and shifting temperatures brought on by climate change.

Reduced crop yields and crop losses can result from floods, droughts, and other extreme weather conditions. Continuous paddy farming without good soil management techniques can eventually result in soil degradation. Frequent problems, including soil erosion and loss of soil fertility, may impact long-term sustainability. To buy seeds, fertilizer, machinery, and other cultivation-related inputs, many rice farmers, mainly smallholders, have restricted access to credit and money [3].

Accessing marketplaces that pay reasonable rates for their produce can be difficult for paddy farmers. Their profitability and income may be affected by price changes. Improvements in production may be hampered by limited access to contemporary agricultural techniques, technologies, and knowledge. Farmers might not be knowledgeable about the most recent approaches to pest management, nutrient management, and other best practices. Post-harvest losses brought on by spoilage, pests, and insufficient processing methods can be caused by poor storage infrastructure and facilities. Despite government support and subsidy programs for agriculture, farmers may find it difficult to navigate them. Implementation can be hampered by inadequate support and delays in the payment of subsidies. Farmers may encounter difficulties in some areas due to issues with land tenure and ownership. Land rights disputes can create uncertainty and prevent investment in agricultural endeavours. In rural areas, younger generations are increasingly leaving agriculture in search of better economic possibilities, which creates a workforce shortage during key crop seasons.

By offering data-driven insights, forecasts, and automation, ML can significantly improve several facets of paddy production. To monitor crop health and find diseases early, ML systems can examine satellite pictures and data from drones. As a result, crop losses are decreased due to prompt intervention and targeted application of pesticides or treatments. ML models can estimate paddy yield with a high degree of accuracy by examining historical data, weather patterns, and other pertinent variables. Farmer decision-making regarding resource allocation, storage, and selling is aided by the information provided [4].

A crop's traits, weather predictions, and real-time soil moisture data can all be processed by ML models to optimize irrigation scheduling. This avoids both over and under irrigation, saving water and promoting crop development. ML can analyze data on soil nutrients and make precise fertilizer application recommendations depending on crop needs. As a result, the environmental impact of excessive fertilizer use is minimized. In pictures taken by cameras or drones, ML models can tell crops from weeds. Herbicide use can be decreased by better-targeting weed control efforts with this information.

For paddy and similar items, ML systems can examine price alterations and market patterns. Agriculturalists can make the best decisions about where and when to sell their produce based on this information. ML can assist farmers in forecasting their labour needs during busy agricultural seasons and can optimize workforce allocation by examining historical labour data and crop requirements. To give farmers real-time advice and insights based on weather forecasts, disease alerts, and other pertinent data, ML can be incorporated into decision support systems. For each farm, ML models may generate individualized recommendations that take into account the soil type, the climate, and previous data, resulting in interventions that are more targeted and successful.

The dataset's outlier and missing value occurrence affect how accurately ML models predict paddy output. Choosing features connected to rice growth requires solid knowledge of paddy cultivation. Collecting real-world data on paddy farming needs a sizable dataset with variables such as

weather information, soil information, historical yields, sensor data, and growth phases [5]. An excessive quantity of redundant or irrelevant features can lead to overfitting and long training durations, which can affect the effectiveness of ML methods. To surmount these threats and to employ ML approaches in paddy crop yield anticipating, this study contributes the following.

- For estimating paddy yield, an HMLCWFS was put forth. The suggested approach selects the most significant features from the PD using five FS approaches: BE, SFS, FI, EFS, and GB approach. The proposed method reconstructs the dataset by applying Poincare's formula to merge the results of each FS technique.
- The suggested approach includes a rice crop management module that provides paddy growers with recommendations to increase paddy productivity in the form of Association rules produced by ML models like DT and RF.
- The suggested approach includes a seed selection recommendation module that takes input from agriculturalists and automatically recommends the paddy verities that are appropriate for field cultivation by tying the information to the real-time PD utilized for training by employing SVM, KNN, and NB methods.

2. Literature Review

Deep Learning (DL) and supervised methods (SVM, KNN, and Adaboost) were proposed by V. Malathi et al. [6] as a framework for classifying pathogen-infected rice leaves. The author tallied the earlier publications connected to their research. A real-time dataset with leaf spots, leaf blight, hispa, and other diseases was gathered by the author. The picture was saved at 224*224 pixels after preprocessing that included rotating, flipping, and other operations.

Using supervised learning methods (SVM, KNN, RF), the image's key characteristics were retrieved and stored in feature maps. The photos of paddy with leaf spot infection are then classified using the CRI-NET-V1 architecture. Using 10-fold cross-validation to evaluate the effectiveness. With 0.997 and 0.994 accuracy, respectively, the Neural Network (NN) and SVM models successfully predict the disease-infected leaves.

An ML model built with (SVR, RBFNN, and ANN) was suggested by Vinson Joshva et al. [7] for estimating the paddy yield in the southern region of Tamilnadu. The author additionally examined 25 pieces of literature and tallied the conclusions relevant to forecasting paddy production. In Tamil Nadu's Cauvery Delta Zone (Thanjavur and Thiruvarur), the author gathered real-time data.

Additionally, the author gathered data on soil factors, including pH range, temperature, and annual rainfall from the southwest and northeast monsoons. The author gathered primary information from 16 paddy fields at Perambalur, which is close to Tiruchirapalli. Multiple linear regression is employed to determine the association between each input variable and the intended class. Then, to estimate the parameter that affects paddy yield, the author used SVN, Generalized RNN (GRNN), Radial Basis Function Network (RBFNN), and Backpropagation NN.

The author demonstrated that GRNN and BPNN performed exceptionally well in forecasting paddy production. The author also compared the rice yearly production throughout all of India's states and showed how crucial alluvial soil, a higher mean temperature, and more rainfall are to increasing paddy yield. A. Suruliandi et al.'s [8] comparison of wrapper FS techniques helped them pinpoint the factors that influence paddy production.

The authors gathered a real-time dataset that included information about the environment and the soil. Recursive Feature Elimination (RFE), Boruta, and Sequential Forward Feature Elimination (SFFE) were employed to preprocess the dataset and identify the key features. Classification techniques such as KNN, NB, DT, SVM, and RF were utilized once the dataset was rebuilt. The findings of the suggested model aid in the process of paddy seed variety selection and crop monitoring for farmers. The author continues by stating that the characteristics chosen by the RFE in conjunction with the Bagging algorithm accurately forecast the paddy yield prediction with 0.8938 accuracy and 0.1062 error rate.

The historical data from Andhra Pradesh, covering the years 2001 to 2020, was gathered by Sangeetham Rohini and S. Narayana Reddy [9]. Additionally, the author built a real-time dataset using remote sensing variables such as evapotranspiration, leaf area index, etc. Using MLR, SVR, and RFR approaches, the author found the characteristics that were significantly associated. To predict the factors that have a significant influence on the yield of paddy, the author uses internal cross-validation and hyperparameter tuning.

To estimate the production of three crops such as rice, potatoes, and wheat, Mahamudul Hasan et al. [10] devised an ML framework called KRR that integrates the RF and KNN methods. The author listed the ten key works that were relevant to their work, along with their benefits and drawbacks. For the two crop seasons from 1969 to 2021, the author gathered real-time datasets from four Bangladeshi states. Kharif and Rabi six ML methods were utilized by the author (SVM, Ridge Regression, RF, NB, and KNN). The author pointed out the elements that increased the production of the chosen crops.

Additionally, the author offers suggestions for keeping an eye on the wheat, potato, and paddy kinds (Aus, Aman,

and Boro). A DL model was suggested by Alexandros Oikonomidis et al. [11]. to forecast soybean production. The author gathered data from four US states between 1980 and 2018. The collection includes 25,345 samples with soil metrics, including dry bulk density, clay, bulk density percentage, etc., and 395 features. Solar radiation, Rainfall, Snow Water Equivalent, and other weather variables.

The author removed 25 characteristics from the dataset by using a 95% thresholding value. With the help of RobustScaler, StandardScalar, and MinMaxScaler, the dataset's values were altered. On the retrieved features, the CNN, XGBoost, and LSTM techniques were employed. R2, MSE, and RMSE were used to gauge performance. When compared to the other methods, the findings revealed that the XGBoost technique predicted the soybean yield with an MSE value of 24.360 and an RMSE of 4.936.

An ML model for crop selection based on soil nutrient characteristics was proposed by S. Bhuvaneswari et al. [12]. To cultivate crops like paddy, sugarcane, and bananas, the author gathered soil straight from farmers' fields and examined it in the M.S. Swaminathan research laboratory, which is located in Thiruvaiyaru village of Thanjavur district. For all three crops, the pH percent of hydrogen in the soil is quite important. The essential characteristics in the soil and water nutrient dataset are extracted using the multi-criteria ranking approach. The dynamic ensembling model's KNN method accurately determines the crop and provides advice to farmers with an accuracy of 91% for rice, 90% for bananas, and 80% for sugarcane.

Using weather, soil, and agricultural factors as inputs, P. Sathya and P. Gnanasekaran [13] employed ML and DL techniques to forecast the paddy output. The author gathered data from 14 agricultural blocks in Tamilnadu's Thanjavur district. The dataset spans the years 2014 to 2021 and includes 12 inputs and 3461 records. The most impacting qualities were found by measuring the correlation of the input parameters using the multiple linear regression approach. The characteristics responsible for paddy yield were discovered after the use of classification models like RF, SVM, and LSTM. When compared to other techniques, the MLR-LSTM method fared well, with an accuracy of 96.6%.

Ruan et al. [14] endeavoured to develop a comprehensive model for anticipating the yield of wheat at a field scale during the growing season. This involved integrating proximal weather and sensing data. The study spanned a decade (2010-2020) and included nine field experiments with varying multi-N rates carried out at five different locations, encompassing diverse wheat varieties. The study utilized nearer sensing detail obtained from a sensor of crop circle during the phase of stem elongation, along with weather details spanning the 30 days leading up to planting until the flowering date. Eleven regression algorithms, comprising both statistical and ML approaches, were employed. This involved the integration of two accretion interludes (aggregated or dis- information) and the incorporation of two feature selection approaches, one based on the coefficient of Pearson co-relation and the other on Recursive Feature Elimination.

In a related study [15], the objective was to illustrate the potential application of specific Hyperspectral Vegetation Indices (HVIs). Using artificial AIs, specifically Ensemble-Bagging (EB) and deep NN, the study aimed to contemplate the yield of soybean and presumably Fungal-Bacterial Induced Odor (FBIO) based on the identified HVIs. Additionally, a hybrid DNN-SPEA2 approach was integrated to estimate optimal HVI values. To remember the most useful HVIs for yield contemplation and FBIO, the approach called feature recursive eliminating wrapper was applied, determining the topmost HVI selections.

3. Background Knowledge

3.1. FS Techniques

Many characteristics are employed to develop a dataset as real-world data is created to build the ML model. How logically the data is organized will determine how well an ML model predicts the target variable. The ML model's accuracy will decline when a dataset's features are added. Additionally, it makes the model biased and adds to the temporal complexity.

One-hot encoding is utilized to transform the categorical variables in this dataset, such as agricultural block, paddy variety, soil kinds, and wind direction, to numerical values. In this feature engineering technique, numerical values with distinct columns were assigned to the categorical categories. FS approaches were applied to prefer the dataset's most influential characteristics. Among the three most widely used strategies namely the embedded method, wrapper, and filter [16]. The primary factors influencing the target variable are identified by this study using the procedures included in the wrapper technique.

3.1.1. BE

The characteristics that are provided to the ML model as input determine how well it learns. By giving the model a significance threshold, BE removes the noisy and undervalued characteristics from the dataset. The BE model receives all of the input characteristics as input, and the model calculates a slope and intercept value to plot a regression line using the input attributes [17].

For each iteration, the attributes were deleted from the attribute list if they did not fit the dataset's regression line. Finally, the dataset still contains the characteristics with significant values less than the significance value. Ordinary

Least Square (OLS) regression is used by the stats model module in the Python library called the scikit learn to construct the regression line, and the equation is provided by,

$$Y = \beta_0 + \sum_{j=1}^{p} \beta_j X_j + \varepsilon \quad (1)$$

In the regression model, Y is the target attribute, the model intercept is portrayed as β_0, X_j, and ε depicts the input, the random error value.

3.1.2. SFS

SFS is an FS technique that starts with no values in the parameters list. The model then computes the significance value for all input features. The feature with the lowest significance value, when compared to all other input variables, is selected and added to the parameters list [3]. The process is repeated until all the attributes which are less than the significance value are set. The Residual Sum of Squares (RSS) is worn to appraise the purpose value of the input variable, and the regression line is built using the equation.

$$RSS = \sum_{i=0}^{n} (\varepsilon_i)^2 \; = \; \sum_{i=0}^{n} \big(y_i - (\alpha + \beta x_i)\big)^2 \quad (2)$$

Here, the x and y are the input variables, α and β are constants, and n is the inspection dataset number.

3.1.3. FI by RF Classifier (FIRFC)

The RF Classifier is employed in FS because it arranges the characteristics in the dataset as nodes of a tree and indicates the priority level of the attributes. The root node in the dataset is chosen as the attribute that has the most significant effect on the target attribute, and the values in the root node are used to segment the dataset [5].

Up till every attribute is divided, the process is repeated. Measuring the impurities in the dataset allowed for the construction of the tree. To separate the dataset into its parts, the RF classifier uses information gain, which is quantified using the equation.

$$Info(D) = \sum_{i=1}^{m} p_i \;\; log2 \; p_i \quad (3)$$

3.1.4. Recursive Feature Elimination (RFE)

RFE is a wrapper-style component selection method where critical features from the dataset are chosen using ML models, including DT, RF, and linear regression. The RF approach is applied in this inquiry to choose the important qualities. Each attribute in the dataset is given a rank by the RF classifier, which also calculates the information gained for each attribute.

The model is rebuilt using the chosen attributes once the least important attribute is eliminated from the list. This procedure keeps on until the dataset has all of the characteristics with the high rank indicated by the RF classifier. The equation for computing the entropy is done using Equation (3), and information gain is given by,

$$\text{Information Gain } (InfoA(D)) = -\sum_{j=1}^{V} \frac{|D_j|}{|D|} \times Info(Dj) \quad (4)$$

$$Gain\ (A) = Info\ (D) - InfoA(D) \quad (5)$$

3.1.5. FS by GB Algorithm

GB predicts the target value using a continuous attribute using both regression and classification approaches. The residual error r1 is calculated using the mean value for the input variable X1 and the target variable y1. The feature matrix (X1, y1) is used to build the DT (Tree1). The feature matrix is rebuilt as (X1, r1) during the following iteration, which employs the residual error r1 as the target variable instead of the original target variable y. By implementing the updated feature matrix and the DT (Tree2), the residual error r2 is determined. Up till all the variables are learned, the trial is imitated. The learning rate (eta) value is multiplied by the prediction of the target variable to arrive at the final prediction of the target variable,

$$y(pred) = y1 + (eta \times r1) + (eta \times r2) + \cdots + (eta \times rn) \quad (6)$$

3.1.6. Poincare's Formula

Poincare's formula is used to union n number of sets, and the general formula is given by,

$$\left|\bigcup_{i=1}^{n} A_i\right| = \sum_{i=1}^{n} |A_i| - \sum_{i \le j \le n}^{n} |A_i \cap A_j| + \cdots + (-1)^r \sum_{1 \le i_1 < \cdots < i_r \le n}^{n} \left|A_{i_1} \cap \ldots \cap A_{i_r}\right| \quad (7)$$

3.2. Supervised ML Techniques

In this kind, the target variable has a label and predetermined values for that label, and the model is schooled using the well-labelled dataset that serves as the input variable. To determine the link between the input and target variable, the model is trained using a mapping function. The dataset is fragmented into datasets of training and testing when using the supervised ML approach. The training set's input characteristics are gathered, and the ML model is then applied to it to predict the target variable. Applying performance measurements to the validation set, the model's performance is evaluated.

3.2.1. DT Induction Technique

Artificial Intelligence expert John Ross Quinlon developed the DT method in 1986. With the smallest dataset, Qunilon developed a technique dubbed ID3 based on Occam's Razor concept. Later, Quninlon refined the technique and presented C4.5, a sophisticated approach for building DTs. The best attribute from the input list is chosen

by this approach, which then divides the dataset based on that attribute. Equations (3), (4), and (5) are used to calculate the attribute selection measure, which is employed to choose the input list's best attribute. Knowledge is extracted for the provided dataset, and association rules are built from the created DT using the (IF-Then) approach.

3.2.2. RF Classifier

The RF approach was developed in 1995 by Tin Kam Ho to get over the problem of overfitting a dataset when building a DT. Leo Breiman and Adele Cutler produced an expansion of the RF and listed it as a symbol in 2006. An RF is a mass of DTs, each of which is made from a portion of the whole dataset. The bootstrap aggregation approach divides a dataset into subgroups, and each subset is constructed as a DT. The RF's final output is selected by majority voting of the classes, and the estimator's average value is computed.

$$\hat{f}_{avg}(X) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^{b}(X) \tag{8}$$

3.2.3. KNN Classifier

A non-parametric classifying technique that classifies the dossier based on the majority vote of the neighbouring class was created in 1951 by Evely Fix and Joseph Hodges. The target class that is closest to the input data's neighbours is determined by calculating the distance. Normalization is performed to increase the KNN model's accuracy when the dataset contains continuous values with variable values. The input variables are collected and organized into an array, and for each array value, the Euclidean distance is computed. Calculating the average of the mean value, which is then utilized as the centroids of a cluster with a fragment of data. Based on the Euclidean distance value of each input variable, a cluster is allocated, and the anticipated target class is determined. The Euclidean horizon betwixt two points (x1, y1) and (x2, y2) is tallied using,

$$dist\left((xi, yi)\right) = \sqrt{(x1 - y1)^2 + (x2 - y2)^2}\ Where\ i = 1,2,3 \ldots . N \tag{9}$$

The smallest Euclidean distance used to assign the target class for the input record X is denoted as $SD_k(x) = \{(y_i^{NN}, c_i^{NN})\}_{i=1}^{k}$. The decision-making function used assigns the input record x into a target class C, which is calculated using,

$$W_c = var \max_{w_j} \sum_{(y_i^{NN}, c_i^{NN}) \in NN_k(x)} \delta\left(c_i^{NN} = w_j\right)$$
$$\text{Where j} = 1,2,3, \ldots \text{M} \tag{10}$$

3.2.4. SVM

SVM is a technique that was proposed by Vladimir Vapnik and his toadies in 1982. It is a versatile ML technique used for both regression and classification undertaking. The dataset with several classes in its target attribute is classified using SVM in a multidimensional space. Support vectors, which are the distance between the two points, are shown in the hyperplane, which has a large margin of error. In this research, let x be an input variable in the PD and the length of the vector is calculated using the Euclidean formula.

$$\|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2} \tag{11}$$

The direction of the input vector in the hyperplane is calculated using,

$$w = \left(\frac{x_1}{\|x\|}, \frac{x_2}{\|x\|}\right) \tag{12}$$

The association between the two input variables, x and y, in the sugarcane dataset, which is plotted in the hyperplane, is calculated with the help of the dot product calculated using the formula.

$$x.y = \|x\|\|y\|\cos(\theta) \tag{13}$$

The hypothesis function that is implemented for predicting the target class of the input variable and deciding in which margin the input variable has to be plotted in the hyperplane is calculated using,

$$h(x_i) = \int_{-1}^{+1} if\ w.x + b \geq 0 \\ if\ w.x + b < 0 \tag{14}$$

The line that is used to sovereign the data plaudits in the hyperplane that has multinomial values is calculated using,

$$k(x, x_i) = sum(x \times x_i) + B(0) \tag{15}$$

3.2.5. Naïve Bayes Classification (NB)

NB technique uses the Bayes theorem as its attribute selection measure. This technique contemplates each trait is responsible for predicting the desired output. Prior and posterior probability is used to indicate the desired outcome, and it is calculated using,

$$P\left(\frac{A}{B}\right) = \frac{P\left(\frac{B}{A}\right)P(A)}{P(B)} \tag{16}$$

Here P (A/B), $P\left(\frac{B}{A}\right)$, P(A), and P(B) portrayed the posterior probability, likelihood, prior probability, and the probability of evidence.

4. Proposed Framework for the HMLCWFS

The districts of Cuddalore and Kallakurichi in Tamilnadu are where the study's data was gathered. A total of 2789 farmers from the agronomic blocks Kallakurichi, Chinna Salem, Panruti, Sangarapuram, Kurinjipadi, and

Cuddalore participated in the survey. From September 15 to January 30 2018, the Late Thaladi season was the time for paddy farming. The dataset includes a total of 45 attributes gathered from the Regional Meteorological Centre in Chennai. It contains meteorological characteristics like monthly temperature, rainfall, instant wind speed, relative humidity, and wind direction. Between the date of sowing and the time of harvest, the climatic parameters were computed monthly. The yield of rice was to be increased by using a novel HMLCWFS model, which was proposed in this study. Figures 1 and 2 below describe the proposed framework's FS method. A dataset is created by preprocessing the unprocessed raw data obtained from farmers, governmental agricultural organizations, and meteorological departments. Using the One-hot encoding technique, the categorical attributes were converted to continuous attributes, and they were then represented as a new column using the Column transform method. The dataset made use of a variety of FS methods, including BE, SFS, FI, RFE, and GB methods.

A new dataset containing the chosen characteristics of the FS process is formed once all FS approaches have been combined, and the critical feature picked by each method is recorded as a subset. The dataset was subjected to supervised ML methods comprising DT, RF, KNN, SVM, and NB methods. Association rules and pattern identification were performed, and the knowledge gleaned was provided to farmers growing paddy for decision-making to increase paddy output.

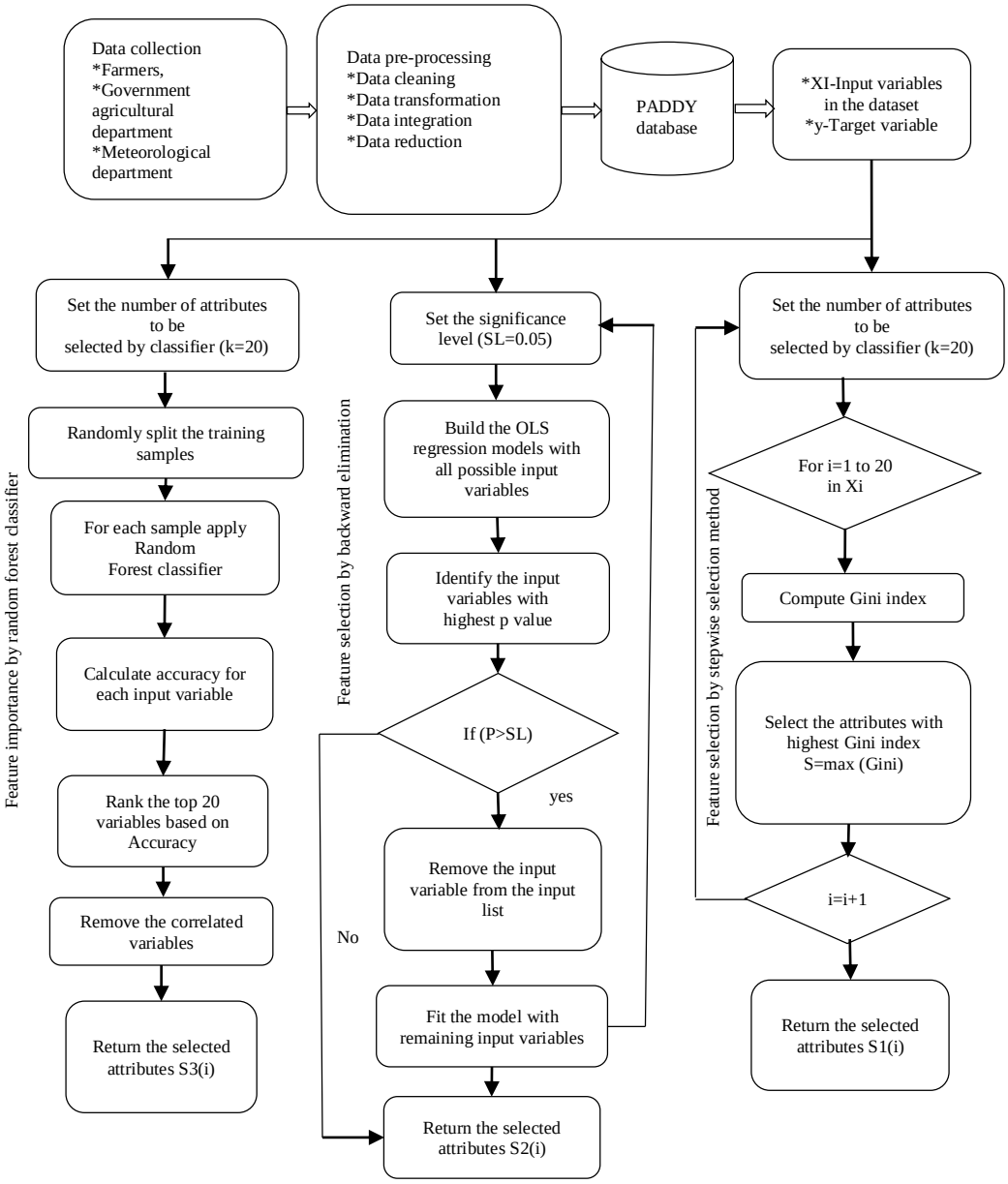

Fig. 1 Flowchart of FS using SFS, BE, and FI by RF classifier

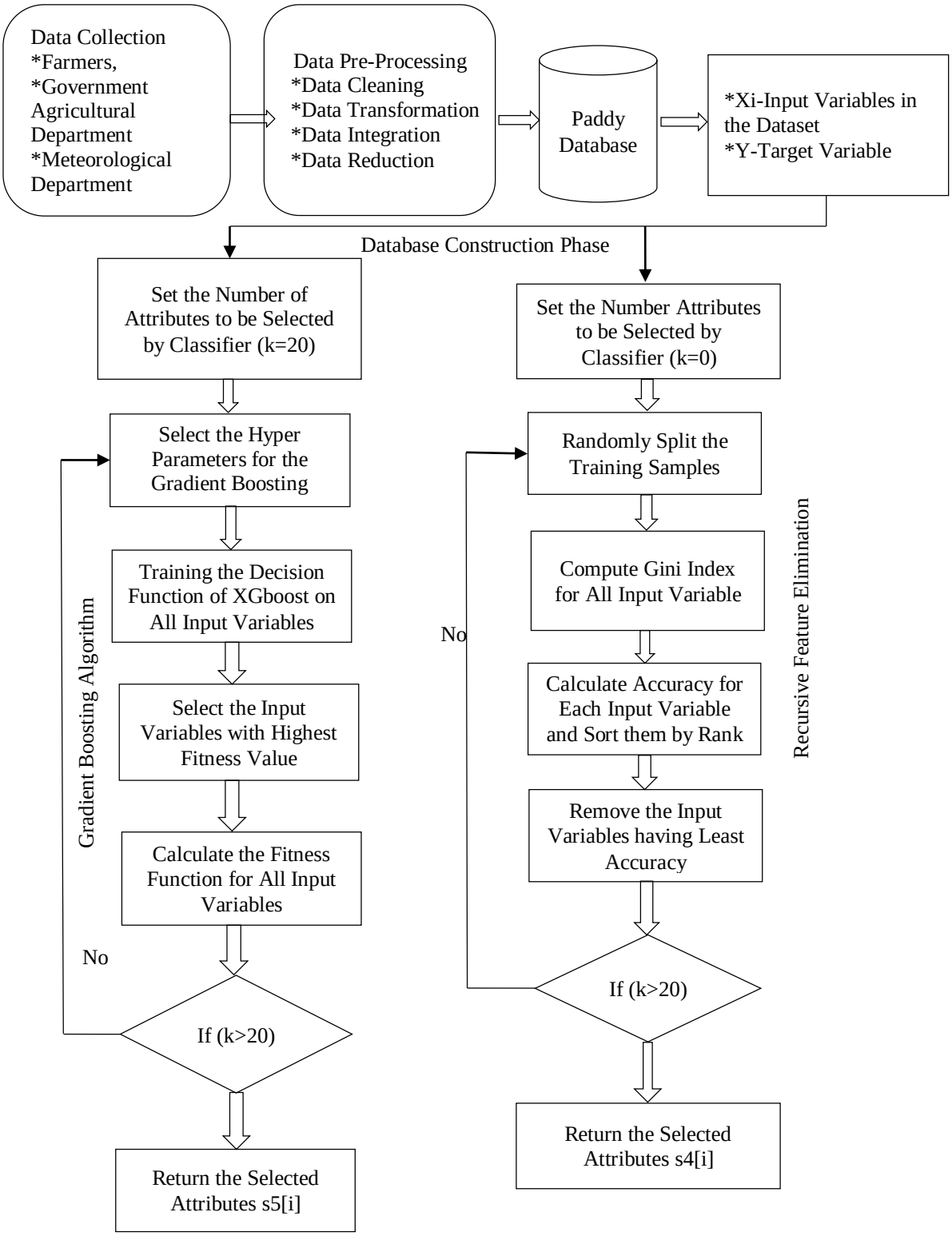

Fig. 2 Flowchart of FS using Recursive Feature Elimination and GB approach

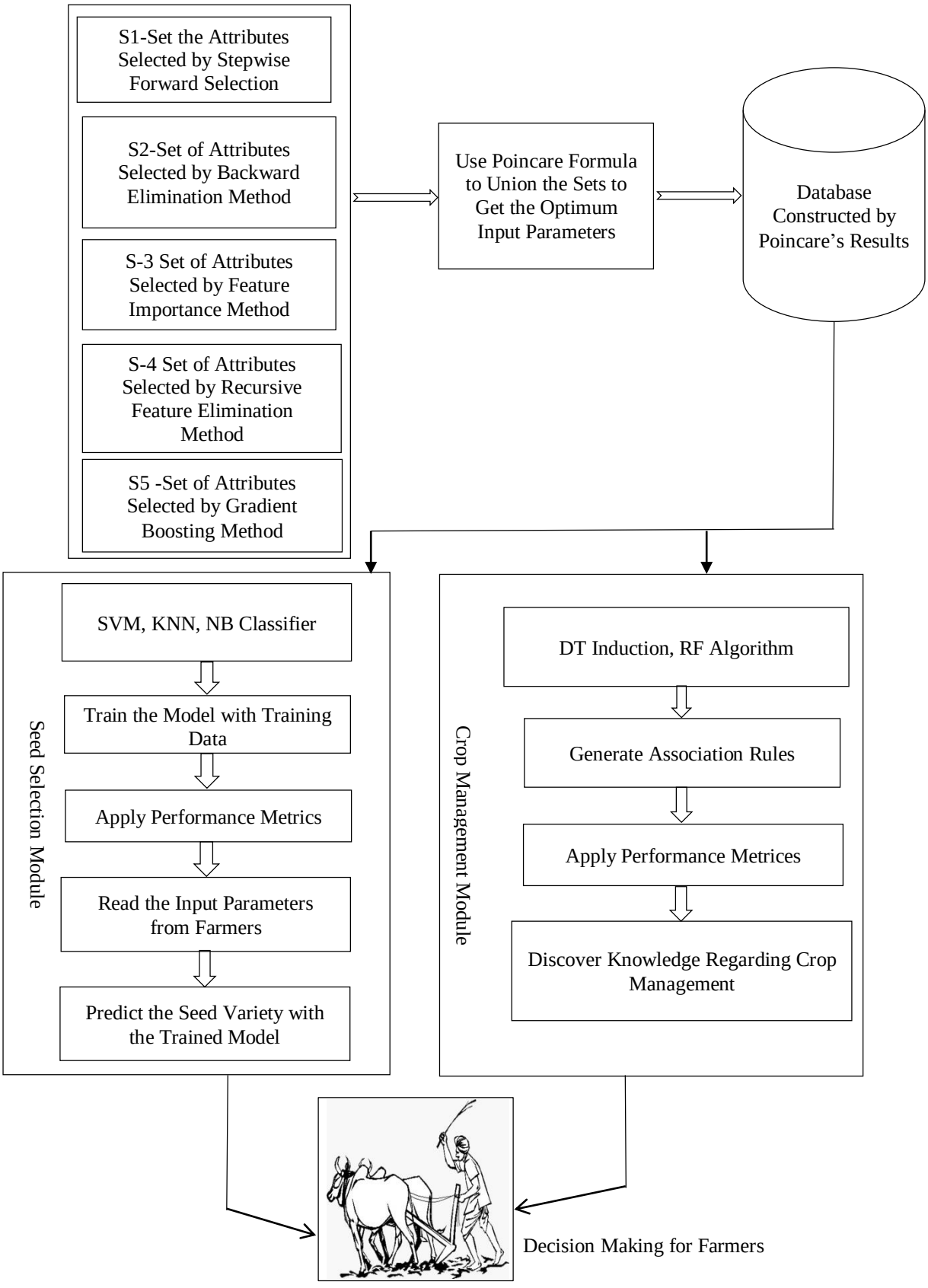

Fig. 3 Architecture of the HMLCWFS approach

5. Experimental Results and Discussion

5.1. BE

The suggested models were put into practice in Python 3.8, with Anaconda Navigator and Spyder 3.8 serving as the code editors. The dataset gathered from the agriculturalists originally contained 45 features, which were then encoded to numerical features using One Hot encoder and Label encoder and transformed to discrete columns by employing Column Transformer from the sklearn package. The final preprocessed dataset has 2,789 records and 70 characteristics, is separated into a testing and training set, and is linearly fit.

The stats model package implements BE using a variety of paddy as the target characteristic. The attribute list is purged of any characteristics with a greater importance level for each cycle. The procedure is repeated until the attribute list contains only attributes with a significance value of less than 0.5. The characteristic with a significance level of zero, shown in Figure 4, indicates that those qualities are the ones with the most influence on the process's outcome.

5.2. SFS

With the help of the RF classifier included in the sklearn.ensemble package, the SFS is carried out. The parameters were set as follows: random_state = 0, n_jobs = 1, and n_estimator = 100, which is used to choose the number of trees. A total of 20 characteristics were selected from the original dataset. The classifier will choose the top 20 attributes out of the total of 70 attributes.

The Gini index is the criterion used to select the characteristic. Each attribute will have the Gini function applied to it by the classifier, which will also calculate the impurity value and give each attribute a rank. The classifier picks the attribute with the greatest rank after eliminating the attribute with the lowest rank. Figure 5 displays the chosen qualities along with their rankings.

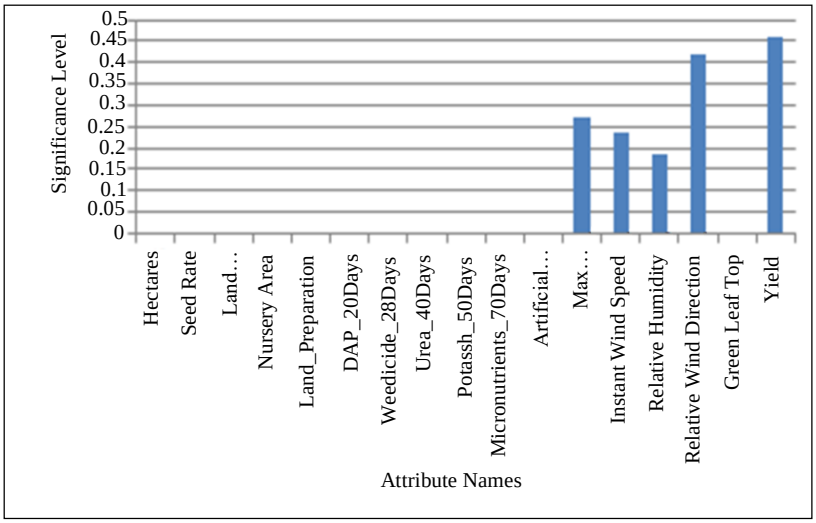

Fig. 4 Attributes selected by BE process

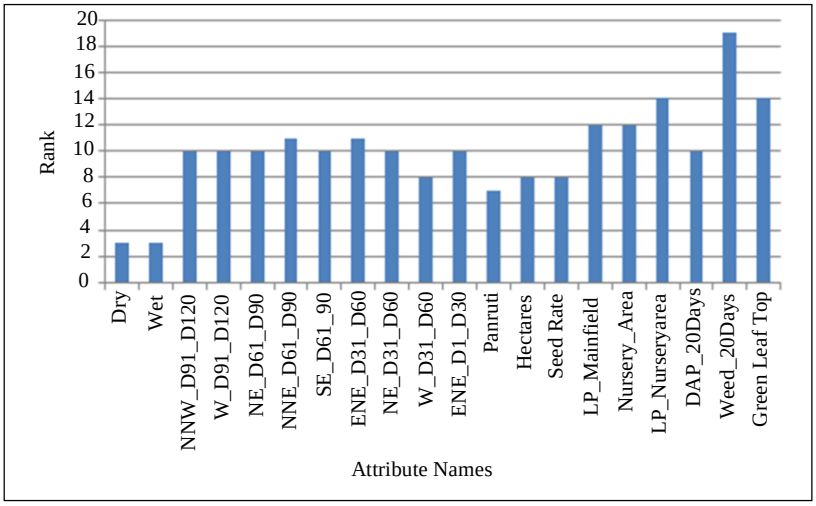

Fig. 5 Attributes selected by SFS

5.3. FI by RF Classifier

This approach is employed with SelectFromModel existing in the package called sklearn.feature_selection. The confines implemented by the RF Classifier are similar to those utilized in the SFS. The sole variance is that the classifier will be routinely choosing the significant attribute numbers from the dataset by computing every attribute's impurity with the gini index function. The attributes chosen by this approach are portrayed in Figure 6.

5.4. Recursive Feature Elimination

SelectFromModel, which is a segment of the sklearn.feature_selection package, is used to implement the FI function. The RF classifier uses the same settings as the SFS. The classifier will automatically choose a certain number of critical characteristics from the dataset by calculating the scum of each attribute using the Gini index function, which is the sole difference. The attributes chosen by this approach are portrayed in Figure 7.

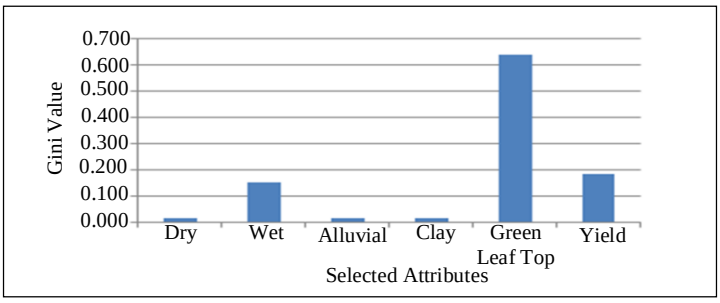

Fig. 6 Chosen features by FI approach

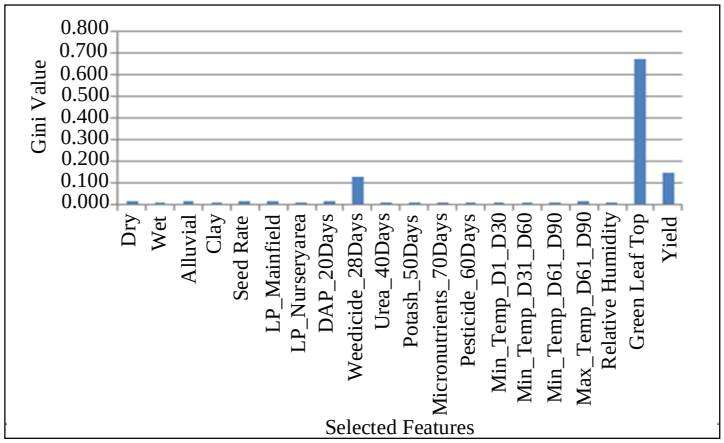

Fig. 7 Selected attributes by Recursive Feature Elimination

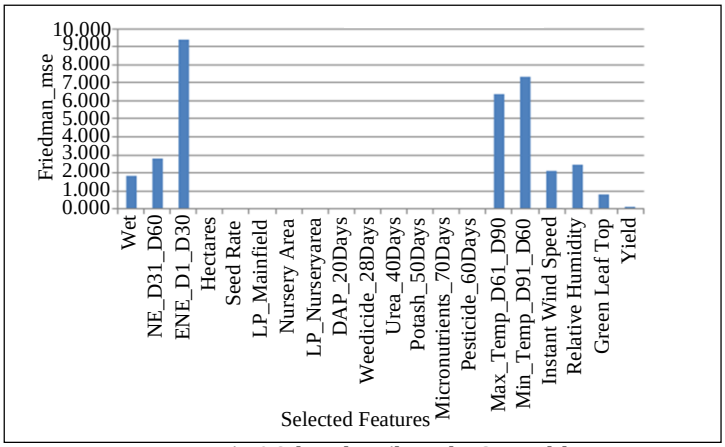

Fig. 8 Selected attributes by GB model

5.5. GB Approach

The GB approach for FS is employed by utilizing the GB Classifier existing in the sklearn.ensemble package. Every attribute of loss is computed by employing the deviance function, and the rate of error was calculated by utilizing the friedman_mse.

The tree was constructed on the multinomial attribute's negative gradient. This classifier chooses the optimum 20 attributes, which are portrayed in Figure 8.

5.6. Poincare's Formula for Integrating the Attributes

As an alternative to utilizing a single FS measure for selecting the best attribute from the dataset provided. This study employs an integration of more FS models and integrates every FS approach's output for getting the most influencing attributes from the dataset. Every FS output is integrated by employing Poincare's Formula or (Inclusion or Exclusion Principle) illustrated as:

$$|\cup_{i=1}^{n} A_i| = \sum_{i=1}^{n}|A_i| - \sum_{i\le j\le n}^{n}|A_i \cap A_j| + \cdots \ldots + (-1)^r \sum_{1\le i_1<\cdots<i_r\le n}^{n}|A_{i_1} \cap \ldots \cap A_{i_r}| \quad (17)$$

The last attribute list attained by integrating the overall five subsets is depicted as,

A_i={'dry', 'wet', 'NNW_D91_D120', 'W_D91_D120', 'NE_D61_D90', 'NNE_D61_D90', 'SE_D61_D90', 'ENE_D31_D60', 'NE_D31_D60', 'W_D31_60', 'ENE_D1_D30', 'Alluvial',' Clay', 'Hectares', 'Seed rate', 'LP_Mainfield', 'Nursery Area', 'LP_nurseryarea',' DAP_20days', 'Weed28D_thiobencarb', 'Urea_40Days', 'Potassh_50Days', 'Micronutrients_70Days', 'Pest_60Day', '51_70Drain', 'Min temp_D1_D30', 'Min temp_D31_D60', 'Min temp_D61_D90', 'Max temp_D61_D90', 'Min temp_D91_D120', 'Inst Wind Speed_D31_D60', 'Inst Wind Speed_D61_D90', 'Relative Humidity_D1_D30', 'Relative Humidity_D31_D60', 'Relative Humidity_D91_D120', 'Green Leaf Top', 'Paddy Yield', 'Paddy Variety'}.

Amongst the overall attributes of 70 provided as input value to the FS procedure, only 38 were attained as this formula's output, as depicted in Ai. Also, the dataset is rebuilt by employing the AI and is enforced for the procedure of ML employment.

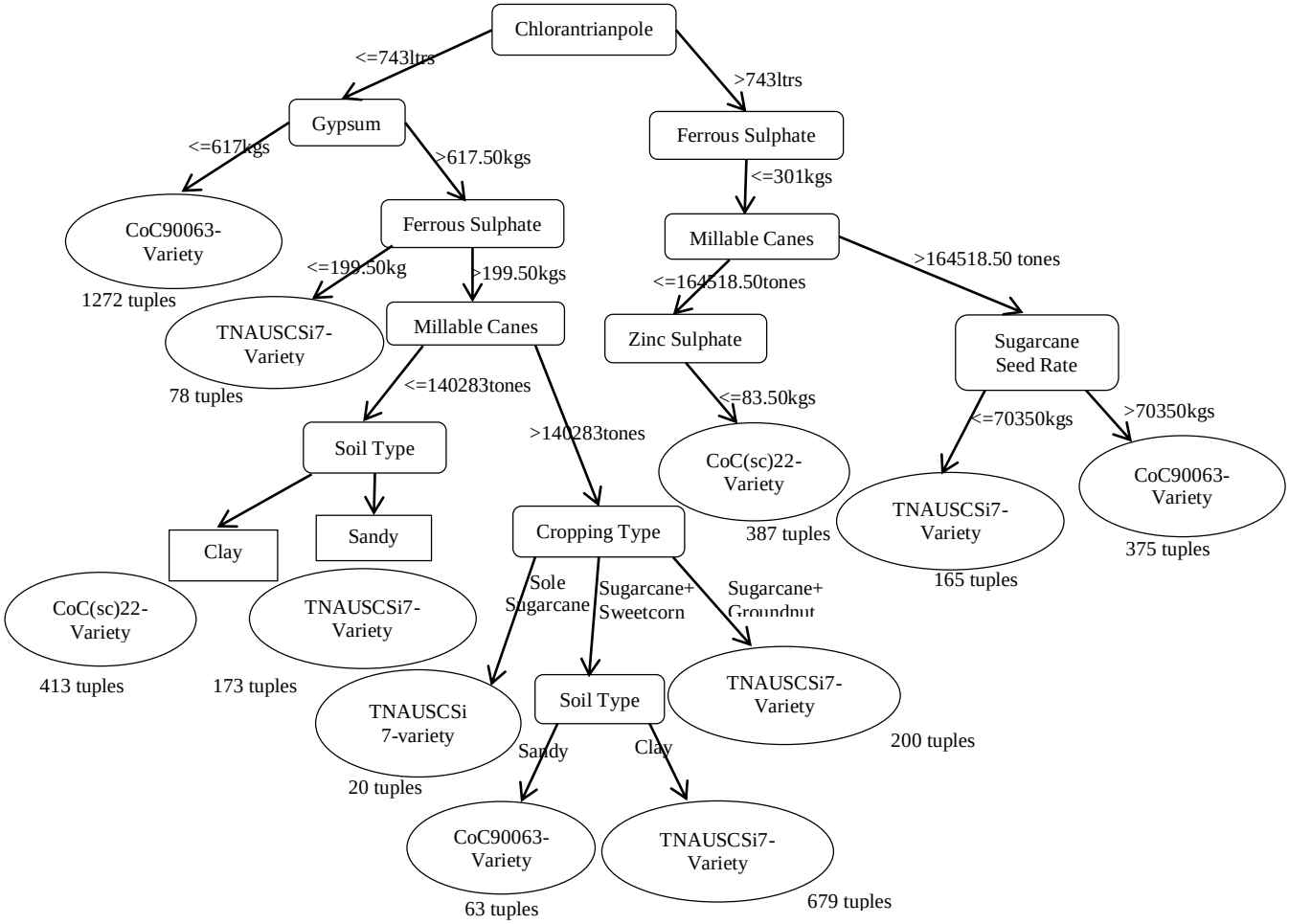

Fig. 9 One of the DTs generated by RF for the PD

5.7. Results for Seed Selection Module

The PD that is rebuilt by Poincare's formula comprises 38 attributes, and among these, the initial 37 attributes are provided as input for constructing the ML approaches. The target attribute to build the ML approach is paddy variety. The initial technique employed in the study is the DT classifier that uses the C4.5 approach for constructing the DT. The dataset comprising 2789 data is segregated into training and testing set comprising 1859 and 930 data. The DT classifier exists in the library of sklearn.

The tree is utilized for building the dataset. The parameters were tuned, and the DT was constructed at various times. The parameters that provide optimal output for the dataset are selected. The criterion, ramdom_state, max_depth, min_samples_leaf, max_features, and min_samples_split is set to entropy, 0, 7, 20, log2, and 50. The constructed DT for the PD is portrayed in Figure 9. The same hyperparameters were used to build the RF classifier, and the only parameter changed is the tree number to be generated, which was set to 5.

The association rules produced from the above DT are depicted in Table 2.

Table 1. Association rules for PD

Generated from the DT for PD R1:If(Chloarantrianpole<=743litres)^(Gypsum<=617kgs) -->(Variety=CoC90063)(1272tuples). R2:If(Chloarantrianpole<=743litres)^(Gypsum>617kgs)^ (Ferrous Sulphate<=199.50kgs)-- >(Variety=TNAUSCSi7)(78tuples). R3:If(Chloarantrianpole<=743litres)^(Gypsum>617kgs)^ (Ferrous Sulphate>199.50kgs)^(Millable Canes<=140283 tones)^(Soil Type=Clay)-- >(Variety=CoC(Sc)22)(413tuples). R4:If(Chloarantrianpole<=743litres)^(Gypsum>617kgs)^ (Ferrous Sulphate>199.50kgs)^(Millable Canes<=140283 tones)^(Soil Type=Sandy)-- >(Variety=CoC(Sc)22)(413tuples).For that 24 rules were extracted from the DT constructed for the sugarcane dataset.

The rules extracted from the DT produced by the C4.5 model existing in Table 1 and Table 2 recommend paddy crop monitoring to the farmers. The minimum temperature of 17°C that occurs between 31 and 60 days after rice seeds are transplanted aids in the development of the plant's reproductive organs and aids in grain filling, success of pollination, and fertilization.

In addition to controlling diseases and pests, the minimum temperature also prevents certain microorganisms from growing in the colder climate. Paddy plant growth is aided by photosynthesis when the highest temperature is not higher than 32°C. This helps prevent heat stress in the paddy, which can cause wilting, decreased nutrient uptake, and slower growth. When the speed of the wind is less than five nautical miles per hour and flows westward for duration of 31 to 60 days, paddy yield is more favourable. This is because high winds promote complete pollination and fertilization while also keeping paddy plants from breaking or bending. The paddy crop grows more quickly when the wind blows from the east and northeast for one to thirty days, then from the west for thirty to sixty days.

An essential factor in paddy production is the use of organic manure during the land preparation phase before cultivation. The paddy plant receives a timely supply of vital nutrients when DAP is applied on day 20, urea is applied on day 40, and potash is applied on day 50. Nitrogen promotes vegetative growth, whereas phosphorous is applied to encourage root development. The use of DAP aided in the development of new shoots from the paddy's main stem.

Table 2. Association rules caused for DTs existing in the RF for PD

Tree1 R1: If(51_70DRain <= 165.70cm)^(Micronutrients_70Days <= 82.50kgs)^(Paddy yield <= 11284kgs)^(Relative Humidity_D31_D60 <= 95.50%)^(Paddy yield <= 11155kgs)→Variety= ponmani(27 tuples) R2: If(51_70DRain <= 165.70cm)^(Micronutrients_70Days <= 82.50kgs)^(Paddy yield <= 11284kgs)^(Relative Humidity_D31_D60 <= 95.50%)^(Paddy yield > 11155kgs)→Variety= CO_43(1 tuples) R3: If(51_70DRain <= 165.70cm)^(Micronutrients_70Days <= 82.50kgs)^(Paddy yield <= 11284kgs)^(Relative Humidity_D31_D60 >95.50kgs)^(Micronutrients_70Days <= 22.50kgs)→Variety= ponmani(16 tuples) Likewise, 27 rules were generated from DT 1 in the RF **Tree 2** R1: If(Max temp_D91_D120 <= 15.25°C)^(Paddy yield <= 31867.5kgs)^(Weed28D_thiobencarb <= 7kgs)^(Green Leaf Top <= 285kgs)^(Urea_40Days <= 67.82kgs)→Variety=CO_43(19 tuples) R2: If(Max temp_D91_D120 <= 15.25°C)^(Paddy yield <= 31867.5kgs)^(Weed28D_thiobencarb <= 7kgs)^(Green Leaf Top <= 285kgs)^(Urea_40Days >= 67.82kgs)→Variety=CO_43(60 tuples) R3: If(Max temp_D91_D120 <= 15.25°C)^(Paddy yield <= 31867.5kgs)^(Weed28D_thiobencarb <= 7kgs)^(Green Leaf Top >= 285kgs)→Variety=Ponmani (338 tuples) Similarly, 25 rules were generated from DT 2 in the RF.

These nutrients enhanced the overall quality of the rice and assisted in filling the grains. Pests that harm rice plants and lower grain yield include rice bugs, leafhoppers, and stem borers. By using insecticides on the sixtieth day, rice

plants can be protected against these pests during the grain-filling stage. Every type of paddy produces a satisfactory yield if there is at least 166 mm of rainfall throughout the 51–70 days period. The results of this study demonstrated that CO_43 supports both clay and alluvial soil, whereas the variety of Ponmani and Delux_Ponni grows best in clay and alluvial soil.

5.8. Results of Seed Selection Module

The KNN classifier is employed with the sklearn neighbours' Python library. The cluster number implemented for splitting the dataset is set to 5, and the Minkowski distance is used for computing the cluster's centroids and the input variable distance. The tuning process of the parameters was achieved, and when the K-value was 5, the classifier gave its optimal output, and the input record was given to the trained model. The SVM is applied with a Support Vector Classifier (SVC) existing in sklearn.svm.

The classifier implements the function of Linear Kernel to compute the values of the kernel matrix to plot the hyperplane input variable. Both testing and training sets are employed for validation, whereas only the training set is utilized for training.

The NB technique is applied with Gaussian NB existing in the library of sklearn.nb. The preceding variance, probability, mean, and the value of absolute additive to the input factors were computed, and the target class for every attribute was recognized by the approach. The seed variety recommended to farmers by the three methods in the seed selection module is provided in Table 3.

Table 3. Seed variety recommended by the various classifiers in the seed recommendation module

Cropping Pattern Received as Input from Farmers	Seed Variety Recommended by the Module	Name of the Model
If(dry nursery=yes)^(direction of wind_Day31_60=west)^(direction of wind_Day1_30=East North East)^(Soil Type=Alluvia)^(Hectares of Land=6)^(Seed Rate=150kgs)^(Nursery Area Mannure=120 cent)^(Land PreparationMannure=6kgs)^(DAP_20days=240kgs)^(Weedicide_28 days=2.4litre)^(Urea_40Days=162.78kgs)^(Potassh_50Days=62.28kgs)^(Micronutirients_70Days=90kgs)^(Pest_60Days=3.6litree)^(51_70Day_Rain=167mm)^(Minimum Temprature_Day1_30=18.5C)^(Minimum Temprature_Day31_60=16C)^(MinimumTemprature_Day61_90=31C)^(Minimum Temprature_Day91_120=16C)^(Instant Wind Speed_D31_60=10knots)^(InstantWindSpeed_D61_90=8knots)^(Relative Humidity_D1_30=72)^(RelativeHumidity_D31_60=78)^(RelativeHumidity_D91_120=85)^(Trash=540bundle)^(Paddy Yield=35028kgs)->Paddy Variety=CO_43.	CO_43	KNN
	Ponmani	SVM
	Delux Ponni	NB

5.9. Plotting the Decision Boundaries and Evaluating the Performance of Each Classifier

The Decision Boundaries (DBs) for the DT and RF classifier shown in Figures 9 and 10 portrayed that the PD is linearly inseparable. The target class variety of paddy containing three values, CO_43, delux_ponni, and Ponmani, was split into each decision region according to the entropy values of the input variables. The classifier plots the input variables in the dataset to one target class for some entropy value. For the same variables, the classifier can assign another decision region to the input variables for another entropy value.

That is, the input variable for one class is repeatedly assigned to another target class, which is clearly shown in the decision boundaries of the DT and RF Classifier. The DB for KNN illustrated in Figure 10 depicted that the PD is linearly inseparable and more convolutional to plot in the decision region. For k value 5, the input variables are plotted clearly in various decision regions, and the target class is easily identified. Figure 13 depicts that the SVM classifier linearly splits the PD and plots clearly in the hyperplane belonging to two regions, and the classifier quickly identifies the target class.

The decision boundaries of the Naïve Bayes classifier shown in Figure 14 tell that the Gaussian NB function used by the classifier determines most of the input records to one particular region, and only a few input records are assigned to another region.

The Gaussian Naïve Bayes classifier also proves that the given PD is linearly separable. The classifier's achievement employed in this study was computed by implementing the performance metrics shown in Figure 15 to Figure 19, and the results are given in Table 4.

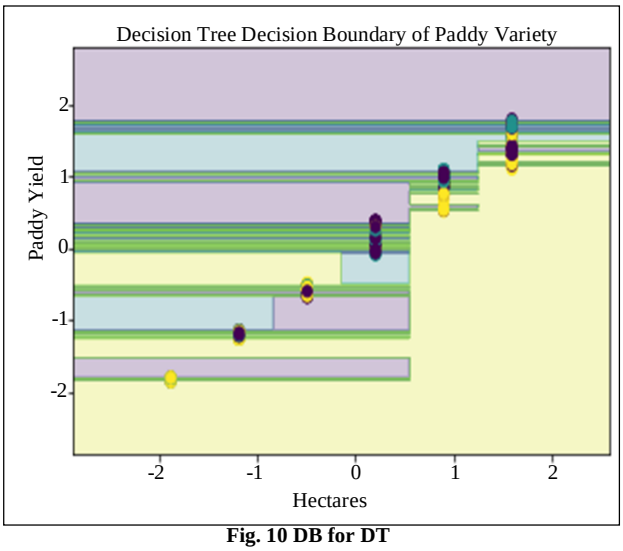

Fig. 10 DB for DT

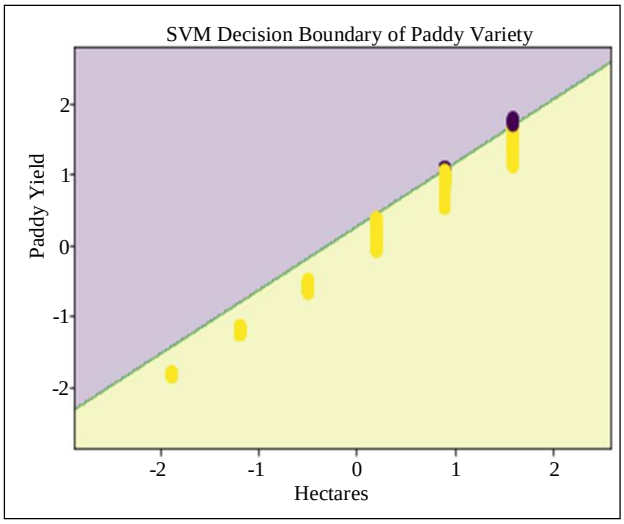

Fig. 13 DB for SVM

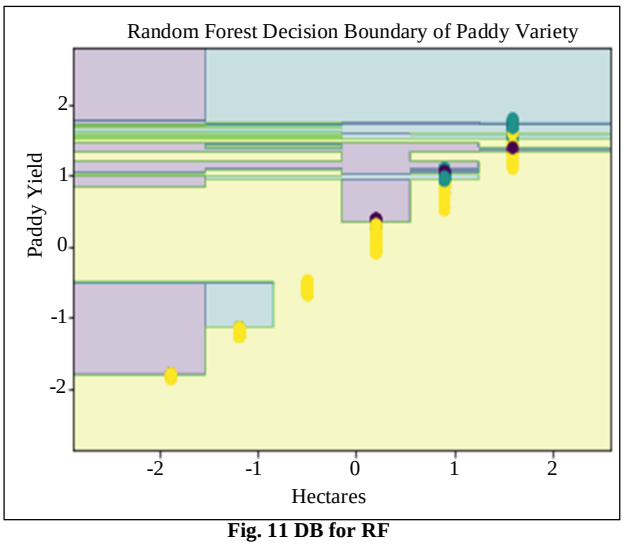

Fig. 11 DB for RF

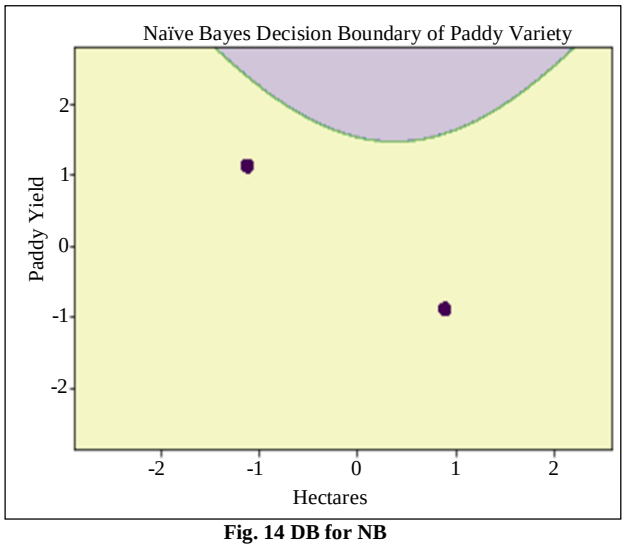

Fig. 14 DB for NB

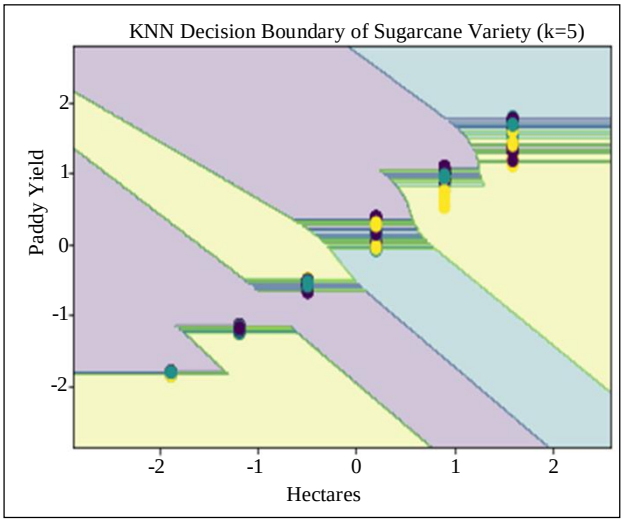

Fig. 12 DB for KNN

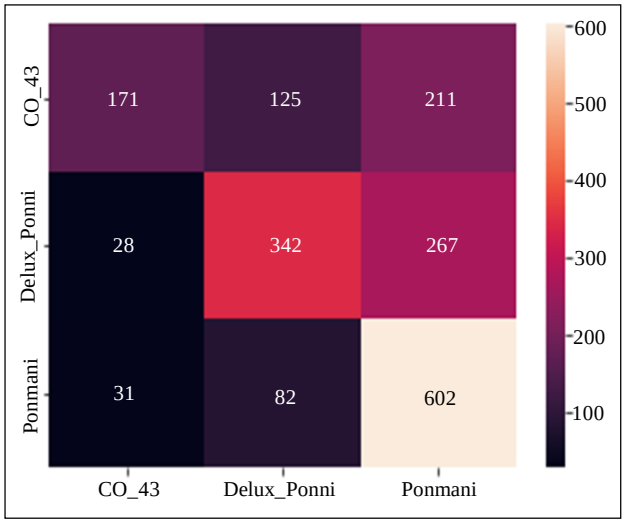

Fig. 15 Confusion matrix for DT

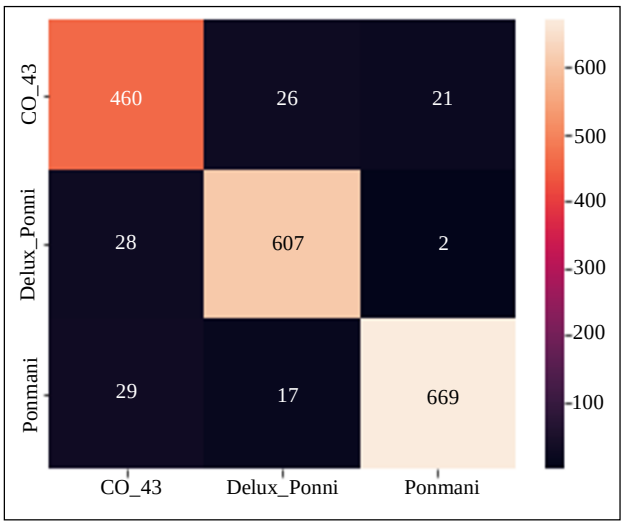

Fig. 16 Confusion matrix for KNN

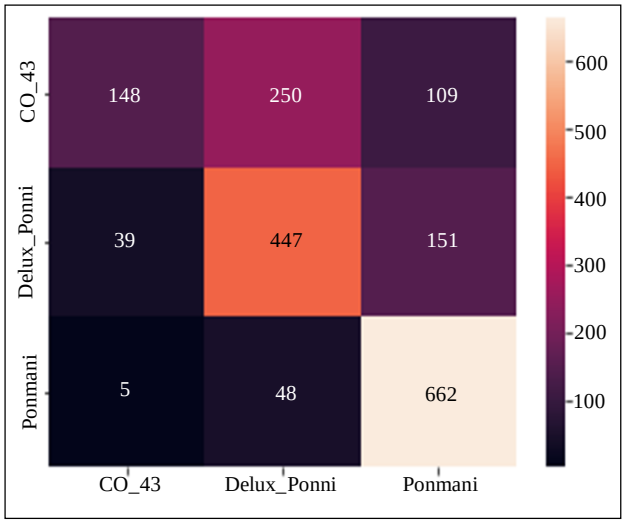

Fig. 17 Confusion matrix for RF

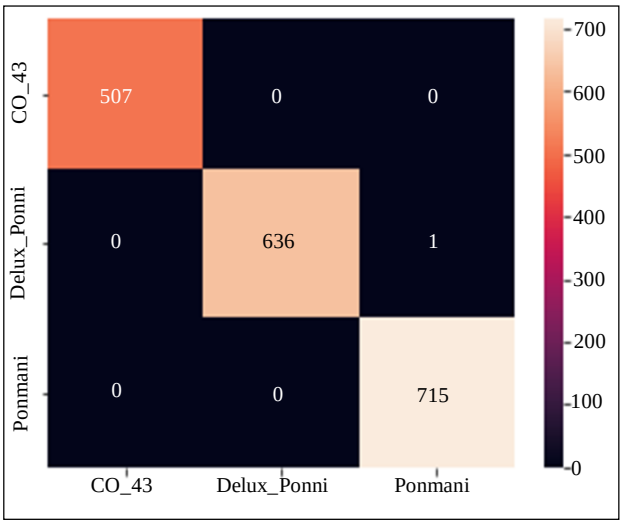

Fig. 18 Confusion matrix for SVM

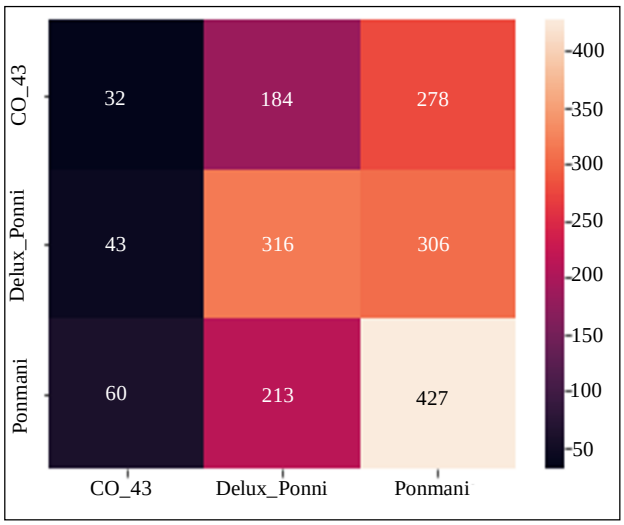

Fig. 19 Confusion matrix for NB

The performance metrics are used to gauge how well ML techniques are applied in classifying the PD. An analysis of the training set results presented in Table 4 reveals that the SVM achieves best in categorizing the data with a 100% accuracy rate and no error rate. With the accuracy and rate of error of 0.96 and 0.04, 0.78 and 0.22, and 0.71 and 0.29 for the training set, the KNN, RF, and DT approaches rank second, third, and fourth places. The Naïve Bayes method performs quite poorly when it comes to classifying the PD when compared to the other five classifiers. The dataset is only 0.60 accurately categorised, and the training set's error rate is 0.40.

Table 4. Performance of the training Set

Training Set	DT	KNN	RF	SVM	NB
Accuracy	0.713	0.96	0.78	1.00	0.60
Error Rate	0.287	0.04	0.22	0.00	0.40
Recall	0.571	0.93	0.64	0.99	0.37
Specificity	0.789	0.97	0.83	1.00	0.69
Precision	0.640	0.93	0.70	0.99	0.36
1-Precision	0.36	0.07	0.30	0.01	0.64
F-Measure	0.57	0.93	0.63	0.99	0.38
False Positive Rate	0.21	0.03	0.17	0.00	0.31
False Negative Rate	0.43	0.07	0.36	0.01	0.63
False Discovery Rate	0.36	0.07	0.30	0.01	0.64
Negative Predicted Value	0.81	0.97	0.85	1.00	0.69

6. Conclusion

The target of the study is to construct an ML approach to surmount the issues encountered while employing ML in the real-time dataset. The proposed HMLCWFS approach to anticipate the paddy yield chooses critical attributes efficiently by using several FS methodologies and resampling the dataset with scarcer attributes. The supervised ML methodologies utilized in this model classify the dataset effectively and identify the aspects that are focused on the cultivation of paddy for enhanced yield. The proposed technique also effectively commends the paddy varieties to be implemented for cultivation by obtaining the input attributes dynamically from the agriculturalist and connecting them with the dataset trained by taking the assistance of the ML methods. Every classifier's achievement was investigated with performance metrics, and every approach's outputs were integrated. The data extracted from the real-time PD was provided to agriculturalists for decision-making during the cultivation of paddy to enhance the yield.

References

[1] Chandraprabha M., and Rajesh Kumar Dhanraj, "Ensemble Deep Learning Algorithm for Forecasting of Rice Crop Yield Based on Soil Nutrition Levels," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 10, no. 4, pp. 1-11, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Md Didarul Islam et al., "Rapid Rice Yield Estimation Using Integrated Remote Sensing and Meteorological Data and Machine Learning," *Remote Sensing*, vol. 15, no. 9, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[3] Baishun Liu et al., "Comparison of Yield Prediction Models and Estimation of the Relative Importance of Main Agronomic Traits Affecting Rice Yield Formation in Saline-Sodic Paddy Fields," *European Journal of Agronomy*, vol. 148, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[4] A. Zhiwei Ye et al., "High-Dimensional Feature Selection Based on Improved Binary Ant Colony Optimization Combined with Hybrid Rice Optimization Algorithm," *International Journal of Intelligent Systems*, vol. 2023, pp. 1-27, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[5] Yue Wang, and Yuechen Li, "Mapping the Ratoon Rice Suitability Region in China Using Random Forest and Recursive Feature Elimination Modeling," *Field Crops Research*, vol. 301, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[6] V. Malathi et al., "Enhancing the Paddy Disease Classification by Using Cross-Validation Strategy for Artificial Neural Network over Baseline Classifiers," *Journal of Sensors*, vol. 2023, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Vinson Joshua, Selwin Mich Priyadharson, and Raju Kannadasan, "Exploration of Machine Learning Approaches for Paddy Yield Prediction in Eastern Part of Tamilnadu," *Agronomy*, vol. 11, no. 10, pp. 1-19, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[8] A. Suruliandi, G. Mariammal, and S.P. Raja, "Crop Prediction Based on Soil and Environmental Characteristics Using Feature Selection Techniques," *Mathematical and Computer Modelling of Dynamical Systems*, vol. 27, no. 1, pp. 117-140, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[9] Sangeetham Rohini, and S. Narayana Reddy, "Machine Learning Based Techniques for Paddy Yield Prediction for the State of Andhra Pradesh," *International Journal of Intelligent Systems and Applications in Engineering*, vol. 1, no. 6S, pp. 753-764, 2023. [Google Scholar] [Publisher Link]

[10] A.S.M. Mahmudul Hasan et al., "A Survey of Deep Learning Techniques for Weed Detection from Images," *Computers and Electronics in Agriculture*, vol. 184, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[11] Alexandros Oikonomidis, Cagatay Catal, and Ayalew Kassahun, "Hybrid Deep Learning-Based Models for Crop Yield Prediction," *Applied Artificial Intelligence*, vol. 36, no. 1, pp. 1933-1950, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[12] Bhuvaneswari Swaminathan, P. Saravanan, and V. Subramaniyaswamy, "Fertilizer Recommendation System for High Crop Yield Based on Prediction Model: A Comparative Analysis," *Advances in Data Science and Computing Technologies*, pp. 1-8, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] P. Sathya, and P. Gnanasekaran, "Ensemble Feature Selection Framework for Paddy Yield Prediction in Cauvery Basin Using Machine Learning Classifiers," *Cogent Engineering*, vol. 10, no. 2, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[14] Guojie Ruan et al., "Improving Wheat Yield Prediction Integrating Proximal Sensing and Weather Data with Machine Learning," *Computers and Electronics in Agriculture*, vol. 195, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[15] Mohsen Yoosefzadeh-Najafabadi, Dan Tulpan, and Milad Eskandari, "Using Hybrid Artificial Intelligence and Evolutionary Optimization Algorithms for Estimating Soybean Yield and Fresh Biomass Using Hyperspectral Vegetation Indices," *Remote Sensing*, vol. 13, no. 13, pp. 1-21, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[16] Anurag Satpathi et al., "Comparative Analysis of Statistical and Machine Learning Techniques for Rice Yield Forecasting for Chhattisgarh, India," *Sustainability*, vol. 15, no. 3, pp. 1-18, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[17] Nilesh Kumar Singh, Shraddha Rawat, and Shweta Gautam, "Weather-Based Rice Crop Yield Forecasting Using Different Regression Techniques & Neural Network Approach for Prayagraj Region," *International Journal of Environment and Climate Change*, vol. 13, no. 10, pp. 2425-2435, 2023. [CrossRef] [Google Scholar] [Publisher Link]