Crop prediction based on soil and environmental characteristics using feature selection techniques

A. Suruliodin\textsuperscript{a}, G. Mariammal\textsuperscript{b} and S.P. Raja\textsuperscript{b}

\textsuperscript{a}Department of Computer Science and Engineering, Manonmaniam Sundaranar University, Tirunelveli, India; \textsuperscript{b}Department of Computer Science and Engineering, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, Tamilnadu, India

**ABSTRACT**

Earlier, crop cultivation was undertaken on the basis of farmers’ hands-on expertise. However, climate change has begun to affect crop yields badly. Consequently, farmers are unable to choose the right crop/s based on soil and environmental factors, and the process of manually predicting the choice of the right crop/s of land has, more often than not, resulted in failure. Accurate crop prediction results in increased crop production. This is where machine learning playing a crucial role in the area of crop prediction. Crop prediction depends on the soil, geographic and climatic attributes. Selecting appropriate attributes for the right crop/s is an intrinsic part of the prediction undertaken by feature selection techniques. In this work, a comparative study of various wrapper feature selection methods are carried out for crop prediction using classification techniques that suggest the suitable crop/s for land. The experimental results show the Recursive Feature Elimination technique with the Adaptive Bagging classifier outperforms the others.

**1. Introduction**

For a nation, one of the most important aspects of its growth revolves around its potential to produce food. For generations, the production of essential food crops has been correlated with agriculture. In reality, however, the rapid pace of population growth has, by far, been the single biggest preoccupation of our society. In doing so, the scope of agriculture has been greatly undermined, particularly in terms of land use and fertility. Given that the area of land under cultivation in this era of urbanization and globalization is unlikely to increase, the focus will have to be on making the most of what there is. In agriculture, crop cultivar prediction is a key factor. Although recent research has opened up statistical information on agriculture, few studies [1–3] have investigated crop prediction based on historical data. However, owing to the unbridled use of fertilizers comprising nitrogen, potassium, and micronutrients, crop cultivation prediction is a challenge. In general, agro-climatic input parameters such as soil texture, rainfall, and temperature influence crop production. Input parameters for agriculture vary...
from region to region, and it is daunting to collect such information over large tracts of land. The vast datasets obtained can be used for crop prediction on a massive scale. Owing to the nature of the problems involved, there is a need to develop new machine learning methods for farming arable land and making the most of narrow land resources. Researchers in agriculture have been testing numerous forecasting methodologies to identify the most suitable crop for specific areas of land.

Predicting suitable crop for cultivation is an essential part of agriculture, with machine learning algorithms playing a major role in such prediction in recent years. In this era of technology and data science, the agricultural sector stands to benefit greatly from properly implemented techniques. Feature selection and classification are critical machine learning techniques [4–7]. Feature selection has to do with selecting the most important attributes from a dataset. It involves picking a subset of appropriate attributes from a larger set of original attributes in terms of a predefined benchmark, such as classification performance or class separability, which plays a significant role in machine learning applications. Three feature selection methods – filter, wrapper, and embedded – [1,8–11] are used in the selection of attributes. Filter methods offer rapid execution, though wrapper methods have a better recognition rate. In this work, wrapper feature selection techniques are used to select the best attributes from the dataset, and classification to predict the most suitable crop for a particular piece of land using the selected attributes. There are three common machine learning techniques: supervised, unsupervised, and reinforcement learning. This work uses supervised learning classification techniques for prediction. The principal contribution of this work is to find the best feature selection technique, with a classification method, to predict the most suitable crop for cultivation, based on factors such as soil and environment.

1.1 Literature survey and justification for the proposed work

Clearly, a farmer is the best decision maker in the selection and cultivation of crops. Today, however, cultivar prediction is done manually in laboratories, and farmers need the help of experts to determine the most suitable crop/s for a specific piece of land. The experts collect soil samples from a particular portion of land and test them in the laboratory, following which they offer suggestions on the ideal crop/s to be raised. Prediction takes time, and selecting the most suitable crop/s is a complex task in agriculture. Manual prediction has largely failed, owing to climatic changes and environmental factors that affect crop cultivation. Accurate predictions of suitable crops for cultivation improve production levels. Crop prediction attributes are defined by multiple factors such as genotype, climate and the interactions between the two. Accurate crop prediction needs a fundamental understanding of the functional relationship between cultivation and interactive factors like the genotype and climate. Further, it requires both detailed datasets and efficient algorithms to examine these relationships. Justified by these facts, machine learning techniques are used in this study to predict the most suitable crop for a specific stretch of land, and this technique is ideal for considering factors like the soil and environmental conditions. A number of related studies are discussed in this review.

Sanmay Das [1] discussed the pros and cons of the filter and wrapper methods, and implemented a new hybrid feature selection approach using the boosting technique. The
experiments were carried out using real-world datasets from the University of California, Irvine (UCI) repository. The results proved that the proposed method is much faster than the wrapper method. Huan Liu and Lei Yu [2] reviewed the existing feature selection algorithm for classification and clustering techniques. Subsequently, an intermediate step on an unifying platform was proposed in their work.

Al Maruf et al. [3] demonstrated the superiority of the gapped k-mer composition and reverse complement features of the k-mer composition over other compositions. The Support Vector Machine (SVM) with the Radial Basis Function (RBF) kernel was used as a classification algorithm. Compared to other approaches, the iRSpot-SF performs considerably better than the Matthews, with a correlation coefficient and sensitivity of 69.41% and 84.57% and it has 84.58% accuracy. Jana Novovicova et al. [6] proposed a feature selection method with no search procedure, and one best suited for multimodal data. Isabelle Guyon and Andre Elisseeff [10], also briefly discussed a feature selection method based on the filter and wrapper approaches and, in addition, defined feature ranking and multivariate feature selection. Jia-You Hsieh et al. [12], in their study, discussed Rice Blast Disease (RBD). The Recursive Feature Elimination (RFE) algorithm with the Auto-Sklearn was used to select key features impacting RBD. The aim of their work was to build a model as a warning mechanism for RBD.

Ron Kohavi and George H. John [13] compared the wrapper and induction methods without feature subset selection, and proceeded thereafter to compare them to Relief, a filter method with feature subset selection. The strengths and weaknesses of the wrapper approach were discussed, and a series of improved designs shown. Isabelle Guyon et al. [14], implemented a Support Vector Machine (SVM) technique based on the Recursive Feature Elimination (RFE) for gene selection. Of the different methods used to select features, the RFE is a newly-developed method that selects features for small sample classification problems.

Marc Sebba and Richard Nock [15], analysed the filter model with information gain and a statistical test. A hybrid model was implemented using a minimum spanning tree that was replaced by the first nearest neighbour. Lei Yu and Huan Liu [16], proposed a correlation filter method termed the fast correlation-based filter. Their technique was verified by two different classification algorithms in terms of real-world data, with and without feature selection.

Petr Somol et al. [17], proposed a flexible hybrid sequential floating search algorithm based on the principles of the filter and wrapper methods. The advantage of the proposed method was its flexibility in terms of a trade-off between the quality of the results versus computational time, as well as enabling the wrapper-based feature selection approach to deal with problems of higher dimensionality. Experiments were carried out using the WAVEFORM dataset from the UCI repository and the SPEECH dataset from British Telecom. Salappa et al. [18], analysed the performance and efficiency of an array of feature selection algorithms with classification methods. Their experimental analysis was carried out on 15 datasets from the UCI repository. The results show that most Feature Selection Algorithms (FSAs) significantly reduce data dimensionality without impacting the performance of the resultant models.

Kursa et al. [19] implemented Boruta, an all-relevant feature selection method which gathers every feature that is critical to the outcome in certain circumstances. By contrast, most traditional feature selection algorithms follow a minimally optimal method in
which they rely on a small subset of features that yield a minimal error on a selected classifier. Marcano Cedeno et al. [20], proposed a feature selection method based on sequential forward selection and the feed forward neural network to find the prediction error as a criterion for selection.

Zahra Karimi et al. [21], implemented a feature ranking method using a hybrid filter feature selection scheme for intrusion detection in a standard dataset. The experimental results show that the proposed technique offers higher accuracy than other methods. Surabhi Chouhan et al. [22], proposed a hybrid combination method of applying the Particle Swarm Optimization – Support Vector Machine (PSO-SVM) to select features from a dataset. Assorted benchmark datasets were tested with this technique.

David Heckmann et al. [23] described that the harnessing of natural variability in photosynthesizing ability as a way to improve yields, through a functional phylogenetic analysis for large-scale genetic screening is a laborious task. The potential for leaf reflectance spectroscopy to estimate photosynthetic efficiency specifications in Brassica oleracea and Zea mays, a C3 and a C4 seed, respectively, were analysed, the findings show that phenotyping leaf reflectance is an effective method to enhance the photosynthetic ability of crops.

Maya Gopal and Bhargavi [24], proposed a wrapper feature selection method featuring Boruta that extracts features from a dataset for crop prediction. The technique improves prediction performance and provides effective predictors. In Boruta, the Z score has the most accurate measure, since it takes into consideration the variability of the mean loss of accuracy among trees in a forest.

Aileen Bahl et al. [25], developed a random forest (RF) model with the RFE for improved prediction accuracy. Maya Gopal and Bhargavi [26] analysed the performance of machine learning (ML) algorithms with a variety of feature selection techniques for crop yield prediction. The results showed that the random forest provides higher accuracy than other ML algorithms. Maya Gopal and Bhargavi [27] proposed sequential forward selection, which is a special sequential feature selection process. It is a greedy search algorithm that attempts to find the ‘optimal’ feature subset by iteratively selecting features based on the performance of the classifiers.

From the literature survey it could be analysed that feature selection and classification are key components of machine learning techniques. This research offers an overview of few techniques that help to pick suitable crops, based on the soil and prevailing environmental conditions. In order to do so, finding key attributes from the dataset is vital for crop prediction, which can be done using feature selection techniques. Feature selection reduces the number of attributes without losing essential information, eliminates redundant data from the dataset, and improves prediction accuracy. In this paper, wrapper feature selection techniques such as the RFE, Boruta, and Sequential Forward Feature Selection (SFFS) are used to select attributes. The SFFS performs best when the optimal subset is small, but since it selects one feature at a time and the loop continues for each attribute selection, it is a time-consuming process. Feature selection and rejection in Boruta are done simultaneously, taking up little time, though it does not undergo through a ranking process. The RFE selects the most accurate attributes based on the ranking method, thereby making better prediction accuracy possible. Classification is an important part of crop prediction as well, since it is used to predict the most suitable crop for a particular piece of land. Classification is used
to predict the class for each record in a dataset. This study uses supervised learning techniques for prediction that handle high-dimensional data, like the k-Nearest Neighbour (kNN) [28], Naive Bayes (NB) [21], Decision Tree (DT) [28], SVM [29], Random Forest (RF) [26], and Bagging [30]. The kNN makes no assumptions about the data, though data scaling is a must. The NB is scalable with large datasets, but does not work well if the training data is not representative of the population. The DT handles both numerical and categorical types of data but takes time to train. The SVM, which is the best algorithm for separable classes, needs a lot of processing time for large datasets. The RF handles huge volumes of data with high-dimensional attributes, though the attributes only work if they have predictive power. Bagging helps with overfitting in the model. It overcomes high variance by using n learners of the same size on the same algorithm. However, it does not help with underfitting in the data. Each algorithm has its own way for the prediction process. So, there is a need to find the best feature selection technique with the classifier for crop prediction process. There are many feature selection techniques with the classifier algorithms that are used to find the important features and also for the prediction process. To the best of our knowledge, as far as we know, there is no research taken to compare the performance of the RFE, Boruta, and SFFS feature selection techniques with the above mentioned classifiers for crop prediction process. Motivated by these facts, this work focuses on finding the best feature selection technique and classifier for predicting the suitable crop/s for a specific land. There is, therefore, a need to execute feature selection with a classifier that can be applied to the crop dataset. To this end, a comparative analysis of different methods is carried out and the performance evaluated. Using these techniques, farmers can easily identify the most suitable crop for their land.

1.2 Organization and outline of the proposed work

Figure 1 depicts the overall process of this work. First, the input data is preprocessed to find the missing values, eliminate redundant data, standardize the dataset, and convert target attributes into factor attributes. Essential attributes are extracted from the preprocessed data using wrapper feature selection techniques. The optimized attributes have classification techniques applied to them, prior to which the dataset is split into training and testing phases. Unknown samples from the training dataset are used to train the classification algorithm to determine the crop that is best suited for cultivation in a specific area of land. The testing dataset is used to predict the crop to be raised, using the trained classifier. Finally, a suitable crop is obtained and the results evaluated using different performance metrics. The analysis reveals the best feature selection technique with an appropriate classification method. The remainder of the paper is organized as follows: Section 2 gives methodology for crop prediction. Section 3 discusses the experimental results followed by conclusion in Section 4.
2. Methodology

2.1 Feature selection

Datasets contain redundant information that harms the classification task. Feature selection is a major task in data analytic research, where datasets have a large number of attributes [26]. Feature selection is chiefly used because (i) it allows the machine learning algorithm to train faster, (ii) it decreases the complexity of the model, and (iii) it makes interpretation easier [26]. It also maximizes the model’s accuracy when choosing the right subset, and prevents overfitting. Three types of feature selection methods are used in the selection of attributes: filter, wrapper, and embedded. Of these, the filter and wrapper methods are primarily used to choose the best attributes. Filter methods are based on performance assessment functions such as distance, knowledge, dependence and consistency, derived directly from the training data. Attribute subsets are selected without resorting to the use of a learning algorithm. Wrapper methods require a predetermined learning algorithm, and use performance as the criterion for assessment. They find attributes that are best suited to the learning algorithm so as to improve performance. The wrapper method typically performs better than the filter method but is more computationally expensive. Nonetheless, wrappers yield feature subsets tailored only for a given learning algorithm, which means that the same subset may not be appropriate for use elsewhere. Embedded methods combine filter and wrapper methods, and use an attribute selection process of their own. This work focuses on wrapper feature selection techniques, such as the following, to choose the best attributes from the original dataset. The techniques are applied to select the best attributes from the soil and environment.

2.1.1 Boruta

The Boruta algorithm is a wrapper feature selection method that is capable of working with any classification method, though it is mainly built around the random forest. It is quick to process and to evaluate the importance of the attributes. Based on the results, the
Z score is evaluated using the mean decrease accuracy and standard deviation. In Boruta, the Z score is the most important measure in this feature selection [24].

Predictor values are shifted and entered with the initial predictors. A random forest classifier is then constructed on the combined datasets. The attribute value is calculated by comparing the initial attributes with the randomized ones [19]. Attributes of greater value are considered more significant than the randomized attributes as well.

The Pseudo code for the Boruta algorithm [11] is given below

\begin{verbatim}
Input : D → inputdataset; RFruns → thenumberofrandomforestruns.
Output : finalSet that contains confirmedSet and rejectedSet

Begin

rejectedSet = ∅

// Let permute be the duplicate the dataset D and shuffle the values in each
column which is stored in the variable shadowAttr

// Let cbind be the combined dataset of the dataset D and shadow Attr that
stored in the variable extended Data

// Let MZSA be the maximum z score value of features

foreach RFruns do

    shadowAttr ← permute(D)

    extendedData ← cbind(D, shadowAttr)

    zScore ← randomForest(extendedData)

    MZSA ← max(zScore(shadowAttr))

    // Let a be each feature

    foreach a ∈ inputdataset do

        if zScore(a) > MZSA then

            hit(a)++

            dif

    end foreach

end foreach

// Let hit be each feature

dif
\end{verbatim}
endfor

endfor

//LettwoSidedEqualityTestbethenullhypothesisisthatimportantvariableisequaltothemaximumimportantoftherandomisedfeatures

for\textbf{each} a \in \text{inputdataset} \textbf{do}

significance(a) \leftarrow \text{twoSidedEqualityTest}(a)

\textbf{if} significance(a) > MZS\textbf{then}

\text{confirmedSet} \leftarrow \text{finalSet} \cup a

\textbf{elseif} significance(a) \leq MZS\textbf{then}

\text{rejectedSet} \leftarrow \text{rejectedSet} \cup a

\textbf{endif}

endfor

\textbf{return} \text{finalSet} \leftarrow \text{rejectedSet} \cup \text{confirmedSet}

\subsection{Sequential Forward Feature Selection (SFFS)}

Numerous recent papers\cite{27,31} on machine learning have proved the usefulness of using feature selection in machine learning in supervised learning functions. These include sequential feature selection (SFS) algorithms, which are strategies that reduce the number of attributes by applying a local search\cite{20}. Sequential Forward Feature Selection (SFFS) and Sequential Backward Feature Selection (SBFS) are the most widely studied versions of these algorithms. This work uses forward selection for feature selection.

The SFFS algorithm is a bottom-up search procedure which starts from an empty set and gradually adds attributes selected by an evaluation function\cite{20}. The best attribute that satisfies a criterion function is included with the current attribute set, that is, one step of the sequential forward selection is undertaken. The algorithm also verifies the possibility of improving the criterion if an attribute is excluded, in which case, the worst attribute (according to the criterion) is eliminated from the set. That is, the algorithm performs one step of the sequential backward selection. The SFFS proceeds dynamically, increasing and decreasing the number of attributes, until the desired 'n' attributes are reached. The SFFS algorithm is based on the Akaike Information Criterion (AIC) value for feature selection\cite{27}.

Sequential feature selection algorithms are a group of greedy algorithms that are used to minimize an original d-dimensional feature space to a subspace of a k-dimensional field, where k < d\cite{32}. 
The pseudo code [32] for the SFFS algorithm

**Input**: \( D \rightarrow \) Dataset, \( D = \{d_1, d_2, \ldots, d_n\} \)

The SFFS algorithm takes the whole \( n \) - dimensional features set as input.

SFFS returns a subset of features \( X_k \); thenumber of selected features \( k \), where

\( k < n, \text{has specified apriori.} \)

**Initialization**: \( X_0 = \emptyset, k = 0 \)

Initializethe algorithm with an empty set \( \emptyset \) (null set) so that \( k = 0 \)

(where \( k \) is the size of the subset).

**Step1**

Selectthethemostsignificant feature

\( x^+ = \arg \max J(x_k + x) \) where \( x \in D - X_k \)

\( k = k + 1 \)

**Step2**

Selectthe least significant feature

\( x^- = \arg \max J(x_k - x) \), where \( x \in X_k \)

if \( J(x_k - x) > J(x_k) \) then

\( X_{k-1} = X_k - x^- \)

\( k = k - 1 \)

GotoStep1

else

GotoStep2

**Termination**: stop when \( k \) equals thenumber of desired features
2.1.3 Recursive Feature Elimination (RFE)

For classification with short training specimens and high dimensionality, feature selection plays a key role in preventing overfitting problems, as well as in optimizing classification. The RFE is a commonly used feature selection method for small sample problems. It fits a model, and weak attributes are removed until the required attributes are attained. The RFE mostly uses the Gini coefficient to rank attributes, based on their importance [25]. It calculates the importance of each attribute as the sum of the number of splits that includes the feature, proportionally to the number of samples it splits. The RFE requires a machine learning model as its input, and the actual number of attributes to be used.

The Pseudo code [33] of the RFE algorithm is given below.

\textbf{Input} : D → Dataset, T → TrainingSet, SetofkfeaturesF = \{f_1, \ldots , f_k\}

\text{RankingMethodRM}(T, F)

\textbf{Output} : FinalrankingR

\textbf{Code}:

Repeat for i in 1 to k

\text{RanksetFusingRM}(T, F)

f* ← lastrankedfeatureinF

R(k - i + 1) ← f*

F ← F - f*

Notwithstanding its performance, the RFE tends to discard ‘weak’ attributes which, when combined with other attributes, could significantly enhance performance. It then reduces recursively the number of functions to be used by rating them using the accuracy of the machine learning model as a metric.

2.2 Classifier

After selecting the most important attributes from the dataset, the classification algorithm is applied on the reduced dataset, using techniques like the k-Nearest Neighbour, Naïve Bayes, Decision Tree, Support Vector Machine, Random Forest, and Bagging. Before applying the classifier algorithm, the optimized dataset is split into training and testing datasets. The classifier algorithm is trained with the training dataset, and the trained classifier is applied in the testing phase. The obtained result is used to predict the crop for a specific area of land. The following subsections describe the classifiers used in this work.
2.2.1 k nearest neighbour
The kNN is a nonparametric [34], supervised learning methodology which uses training sets to classify data points into specific categories. Information is collected from all educational cases, as well as the correlations in basic classifications, based on a new case. The training dataset is checked for the greatest number of previous cases (neighbour) k, and new instances (x) estimated by summing up the output attributes for the k cases.

2.2.2 Naive bayes
One of the classifiers in a family of elegant probabilistic categorization methods in machine learning is the Naive Bayes [35], centred on the freedom attributes in the Bayes’ theorem. Every class label is predicted by the likelihood of a given instance. Naive Bayes classifiers conclude that, provided the class attribute, the value of a particular quality is independent of the value of any other quality.

2.2.3 Decision tree
The Decision Tree is a predictive model that works by testing conditions at each tree level, and moving down the tree where different decisions are identified [34]. The situation depends on the application of machine learning and the decision-making outcome. Decision tree algorithms include the C4.5, CART, and ID3.

2.2.4 Support vector machine
The SVM breaks data into decision surfaces, which further divide the data into two hyperplane groups [36]. Training points specify the vector which supports the hyperplane. Presumably, because of greater margins, a hyperplane with the greatest distance to the nearest learning data point usually has better margins and smaller errors, with a high classifier generalization.

2.2.5 Random forest
The Random Forest creates and merges several decision trees to get the most reliable forecast [37]. The RF searches for the most significant parameter of all while dividing each node, following which it search for the best from the subset of random attributes.

2.2.6 Bagging
Bagging is also known as Bootstrap Aggregating. Bagging is an ensemble machine learning method that combines weak learners and learns them parallelly, that is, independent of each other [30]. For the prediction process, the samples are taken from the training dataset to train the classifier [38]. Further, Bagging takes votes for each sample to improve the performance of prediction. A bagging algorithm does not allow weights to be recalculated; thus, there is no need to change the weight update equation or to modify the algorithm’s calculations. In this work Adaptive Bagging (AdaBag) classifier is used for the prediction process. It is very efficient when the bias of the predictor is larger than the variance.

The basic pseudo code for the bagging algorithm [38]
**Input**: $D \rightarrow$ Dataset, $T \rightarrow$ Training set of examples of size $N$, $s \rightarrow$ number of bootstraps samples, $LA \rightarrow$ Learning Algorithm

**Output**: $C$'bagging ensemble with $s$ component classifiers

```
for i = 1 .. s do
    $S_i \leftarrow$ bootstrapsamplefrom $T$
    Generate classifier $C_i \leftarrow LA(S_i)$
endfor
```

Predicting class label for new instance $x$:

$$C^*(x) = \arg\max_y \sum_{i=1}^{s} [C_i(x) = y]$$

### 2.3 Crop prediction procedure

The basic crop prediction procedure is given below. Soil parameters and environmental conditions are given as input and the predicted crop as output.

1. Step 1: Crop dataset is given as input, and the set of data imported.
2. Step 2: The attributes used in the set of data are transformed into a particular range, bringing the set of data into a consistent state, thus avoiding anomalies. Any missing values are removed and normalization used to standardize the data. Redundancy is minimized once, the dataset is structured and also it helps to make the efficient data for the prediction processing.
3. Step 3: The feature selection technique is applied on the preprocessed data to select the most important attributes from the dataset to create a reduced dataset.
4. Step 4: The reduced dataset is split in order to be used in the training and testing phases.
5. Step 5: First, 70% of the samples from the reduced dataset are taken as training samples.
6. Step 6: The classification algorithm is applied to the training samples.
7. Step 7: The classification algorithm is trained with the entire training dataset to predict a suitable crop.
8. Step 8: Of the samples, 30% are taken from the reduced dataset as testing samples.
9. Step 9: The trained classifier is applied to the testing samples to predict the most suitable crop for cultivation in a particular piece of land.
10. Step 10: The target label for new instances is found by the trained classifier so as to identify a suitable crop.
11. Step 11: Finally, a suitable crop for cultivation is recommended by the results.
The basic pseudo code for this work is given below.

**Input**: DatasetD with E classes

**Begin**

\[ Y \leftarrow \text{Replacing missing values using preprocess(D)} \]

\[ X \leftarrow \text{select the important features(Y)} \]

**Train** \[\leftarrow\] samples 70% from \( X \)

**Test** \[\leftarrow\] remainingsamples from \( X \)

\( A \) : Number of Samples \( [\text{Train}] \)

\( B \) : Labels, where \{1, \ldots, N\}

\( M \) : ClassificationModel

ForeachLabel in \{1, \ldots, N\}

Construct a new label vector for E classes

Apply \( A \) to \( M \), to obtain the prediction

Test the model using ‘Test’

Compute Confusion Matrix for ‘Test’ on the basis of prediction

### 3. Experimental result analysis & discussions

#### 3.1 Dataset description

This work utilized an agricultural dataset that chiefly included soil characteristics and environmental factors, collected from the website: [www.tnau.ac.in](http://www.tnau.ac.in) and the Agricultural Department of Sankarankovil Taluk, Tenkasi District, Tamil Nadu, India. The dataset contains 1000 instances, 16 attributes, where 12 attributes are soil characteristics and the remaining 4 environmental characteristics, respectively. The target class is the multiclass representation with 9 classes. The attributes are collected from villages around Sankarankovil.

Table 1 shows a description of the soil and environmental attributes used for crop prediction.
### Table 1. Description of the crop dataset.

| No. | Attributes                      | Description                                                                 |
|-----|---------------------------------|-----------------------------------------------------------------------------|
| 1.  | pH (potential of Hydrogen)      | pH is the main factor for farming.                                          |
| 2.  | EC (Electrical Conductivity)    | EC is the numerical parameter that used to measure the salt level in soil and it affects the crop productivity. If EC is 0.01 then the soil is considered for crop cultivation. |
| 3.  | OC (Organic Carbon)             | OC enters the soil through the decomposition of plant and animal residues, root exudates, living and dead microorganisms, and soil biota. |
| 4.  | N (Nitrogen)                    | Nitrogen is a key element in plant growth.                                  |
| 5.  | P (Phosphorus)                  | Phosphorus helps transfer energy from sunlight to plants, stimulates early root and plant growth, and hastens maturity. |
| 6.  | K (Potassium)                   | Potassium increases vigour and disease resistance of plants, helps form and move starches, sugars and oils in plants, and can improve fruit quality. |
| 7.  | S (Sulphur)                     | Sulphur is a constituent of amino acids in plant proteins and is involved in energy-producing processes in plants. |
| 8.  | Z (Zinc)                        | Zinc helps in the production of a plant hormone responsible for stem elongation and leaf expansion. |
| 9.  | B (Boron)                       | Boron helps with the formation of cell walls in rapidly growing tissue. Deficiency reduces the uptake of calcium and inhibits the plant’s ability to use it. |
| 10. | Fe (Iron)                       | Iron is a constituent of many compounds that regulate and promote growth.   |
| 11. | Mn (Manganese)                  | Manganese helps with photosynthesis.                                        |
| 12. | Cu (Copper)                     | Copper is an essential constituent of enzymes in plants.                   |
| 13. | Texture                         | It has major influence on crop growth. It influences aeration, water movement etc… |
| 14. | Season                          | Season is the challenging factor for crop growth.                          |
| 15. | Rainfall                        | Rainfall has the great impact on crop growth. Excessive and insufficient rainfall affects the yield. |
| 16. | Average Temperature             | It is important for growth and development.                                |

### 3.2 Performance metrics

Performance metrics provide hard data to support evaluation. The results are evaluated using distinctive performance metrics like Accuracy, Error Rate, Kappa, Precision, Recall, F1 Score, Mean Absolute Error, and Log Loss to recommend the best feature selection method for prediction.

#### Accuracy

Accuracy is the number of correct predictions divided by the total number of predictions.

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (1)
\]

where, TP- True Positive; TN- True Negative; FP- False Positive; FN- False Negative.

#### Kappa

It is a measure of agreement between two individuals. Kappa is always less than or equal to 1. Value 1 implies perfect agreement and less than 1 implies less than the perfect agreement.

\[
\text{Kappa} = \frac{P_{\text{agree}} - P_{\text{chance}}}{1 - P_{\text{chance}}} \quad (2)
\]
where, $P_{\text{agree}}$ – Probability of agreement; $P_{\text{Chance}}$ – Probability of agreement due to chance.

**Precision**

The present of predicted positive that is actually positive. It also called as Positive Predicted Value (PPV).

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3).$$

**Recall**

The proportion of positive results out of the number of samples which were actually positive. It is also known as Sensitivity.

$$\text{Recall} = \frac{TP}{TP + FN} \quad (4).$$

**Specificity**

The proportion of negative results out of the number of samples which were actually negative.

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (5).$$

**F1 Score**

F1 Score is the harmonic mean of Precision or PPV and Recall or True Positive Rate (TPR).

$$\text{F1Score} = 2 \frac{\text{PPV} \cdot \text{TPR}}{\text{PPV} + \text{TPR}} \quad (6).$$

**Mean absolute error (MAE)**

MAE is the error metric. It is defined as the average of difference between original values and predicted values.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |Y_i - \hat{Y}_i| \quad (7).$$

Where, $N$ – Number of samples; $Y_i$ – Original values and $\hat{Y}_i$ – Predicted values.
**Log loss**

It is defined on probability estimates and measures the performance of a model. It is also known as cross entropy loss.

\[
\text{LogLoss} = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{E} y_{ij} \log(p_{ij})
\]

where \(N\) – Number of samples; \(E\) – Number of classes; \(y_{ij}\) – whether the sample \(i\) belongs to the class \(j\) or not; \(p_{ij}\) – the probability of sample \(i\) belonging to the class \(j\).

### 3.3 Results and discussion

Attributes are selected, using the feature selection method, to find accurate soil and environmental characteristics for predicting a suitable crop for improved cultivation. Classification methods are used with feature selection techniques to find the most suitable crop for a particular stretch of land. The techniques are evaluated thereafter, using parameters such as Attribute Selection, Accuracy, and Error Rate.

#### 3.3.1 Performance comparison of feature selection with classifier based on soil characteristics

Table 2 shows a performance evaluation of feature selection methods with classification techniques, based only on soil characteristics such as the Ph, EC, N, P, and K, among others, as described in Table 1.

From the 12 attributes, the RFE, Boruta and SFFS select 9, 12, and 10 attributes, respectively for the crop prediction. Classifier techniques are applied to find the most suitable crop/s, based on soil conditions. Table 2 shows that the RFE selects the most accurate attributes of all the techniques. Further, the RFE with the bagging classifier offers

| Feature Method          | Selection Method | Attribute Selected | Accuracy | Error Rate |
|-------------------------|------------------|--------------------|----------|------------|
| **Before Reduction**    |                  |                    |          |            |
| **After Reduction**     |                  |                    |          |            |
| Boruta                  | kNN              | 12                 | 0.5625   | 0.4375     |
| Naïve Bayes             |                  | 12                 | 0.6812   | 0.3188     |
| Decision Tree           |                  |                    | 0.7125   | 0.2875     |
| SVM                     |                  |                    | 0.7750   | 0.225      |
| Random Forest           |                  |                    | 0.8375   | 0.1625     |
| Bagging                 |                  |                    | 0.8875   | 0.1125     |
| Sequential Forward      | kNN              | 12                 | 0.5625   | 0.4375     |
| Feature Selection       |                  | 10                 | 0.5625   | 0.4311     |
| Naïve Bayes             |                  | 12                 | 0.6812   | 0.3188     |
| Decision Tree           |                  |                    | 0.7125   | 0.2875     |
| SVM                     |                  |                    | 0.7750   | 0.225      |
| Random Forest           |                  |                    | 0.8375   | 0.1625     |
| Bagging                 |                  |                    | 0.8875   | 0.1125     |
| Recursive Feature       | kNN              | 12                 | 0.5625   | 0.4375     |
| Elimination             |                  | 9                  | 0.5625   | 0.4375     |
| Naïve Bayes             |                  | 12                 | 0.6812   | 0.3188     |
| Decision Tree           |                  |                    | 0.7125   | 0.2875     |
| SVM                     |                  |                    | 0.7750   | 0.225      |
| Random Forest           |                  |                    | 0.8375   | 0.1625     |
| Bagging                 |                  |                    | 0.8875   | **0.8938** |

| Attribute Selected | Accuracy | Error Rate |
|--------------------|----------|------------|
| Boruta             |          |            |
| kNN                | 0.5625   | 0.4375     |
| Naïve Bayes        | 0.6812   | 0.3188     |
| Decision Tree      | 0.7125   | 0.2875     |
| SVM                | 0.7750   | 0.225      |
| Random Forest      | 0.8375   | 0.1625     |
| Bagging            | 0.8875   | 0.1125     |
| Sequential Forward | 0.5625   | 0.4311     |
| Feature Selection  | 0.5625   | 0.4375     |
| kNN                | 0.6812   | 0.3188     |
| Naïve Bayes        | 0.7125   | 0.2875     |
| Decision Tree      | 0.7750   | 0.225      |
| SVM                | 0.8375   | 0.1625     |
| Random Forest      | 0.8875   | 0.1125     |
| Bagging            | 0.8938   | **0.1062** |
| Recursive Feature  | 0.5625   | 0.4375     |
| Elimination        | 0.5625   | 0.4375     |
| kNN                | 0.6812   | 0.3188     |
| Naïve Bayes        | 0.7125   | 0.2875     |
| Decision Tree      | 0.7750   | 0.225      |
| SVM                | 0.8375   | 0.1625     |
| Random Forest      | 0.8875   | **0.1062** |

**Table 2.** Performance comparison of feature selection methods based on soil characteristics.
better crop prediction accuracy of 89%, compared with other soil-characteristic-based methods.

### 3.3.2 Performance comparison of feature selection with classifier based on environmental conditions

Table 3 represents a performance analysis of feature selection with classification techniques, based on environmental factors alone, such as textures, seasons, rainfall and average temperature.

From the remaining 4 attributes, the RFE selects 2 attributes, while Boruta and SFFS select 3 attributes each for the prediction process. Classifier techniques are then applied to find the most suitable crop/s, based on environmental characteristics. Table 3 shows that the RFE selects the most accurate attributes of all the techniques. In addition, the RFE with bagging offers better crop prediction accuracy of 56%, compared with other environmental-characteristic-based methods.

### 3.3.3 Performance comparison of Feature selection with Classifier based on Soil and Environmental Characteristics

Table 4 represents a performance analysis of feature selection with classification techniques, based on the soil and environmental conditions such as N, P, K, rainfall, average temperature, etc . . .

From the 16 attributes in all, the RFE, Boruta and SFFS select 14, 15, and 13 attributes, respectively for predicting the suitable crop. Classifier techniques are then applied to find the most suitable crop/s, based on both soil and environmental characteristics. Table 4 shows that the RFE selects the most accurate attributes of all the techniques. In addition, the RFE with bagging offers better crop prediction accuracy of 92%, compared to other methods based on both soil and environmental characteristics.

In this work, the dataset contains only 16 attributes. Although, the dataset is reducible dataset still it is needed to reduce the attributes for improving the prediction level. The given feature selection techniques are selecting the relevant features from the dataset and

**Table 3.** Performance comparison of feature selection methods based on environmental conditions.

| Feature Method | Selection Method | Classification Method | Attribute Selected | Accuracy Before Reduction | Accuracy After Reduction | Error Rate Before Reduction | Error Rate After Reduction |
|----------------|------------------|-----------------------|--------------------|---------------------------|--------------------------|-----------------------------|---------------------------|
| Boruta         | kNN              | 4                     | 0.3500             | 0.3510                    | 0.6500                   | 0.6490                      |
|                | Naive Bayes      | 4                     | 0.3512             | 0.3523                    | 0.6488                   | 0.6477                      |
|                | Decision Tree    | 4                     | 0.4435             | 0.4493                    | 0.5565                   | 0.5507                      |
|                | SVM              | 4                     | 0.4500             | 0.4557                    | 0.5500                   | 0.5443                      |
|                | Random Forest    | 4                     | 0.4625             | 0.4720                    | 0.5375                   | 0.5280                      |
|                | Bagging          | 4                     | 0.5400             | 0.5500                    | 0.4600                   | 0.4500                      |
| Sequential Forward Feature Selection | kNN | 4 | 0.35 | 0.3510 | 0.6500 | 0.6490 |
|                | Naive Bayes      | 4 | 0.3512 | 0.3523 | 0.6488 | 0.6477 |
|                | Decision Tree    | 4 | 0.4435 | 0.4493 | 0.5565 | 0.5507 |
|                | SVM              | 4 | 0.45 | 0.4557 | 0.5500 | 0.5443 |
|                | Random Forest    | 4 | 0.4625 | 0.4720 | 0.5375 | 0.5280 |
|                | Bagging          | 4 | 0.54 | 0.5500 | 0.4600 | 0.4500 |
| Recursive Feature Elimination | kNN | 4 | 0.3500 | 0.3517 | 0.6500 | 0.6483 |
|                | Naive Bayes      | 4 | 0.3512 | 0.3580 | 0.6488 | 0.6420 |
|                | Decision Tree    | 4 | 0.4435 | 0.45 | 0.5565 | 0.5500 |
|                | SVM              | 4 | 0.4500 | 0.4583 | 0.5500 | 0.5417 |
|                | Random Forest    | 4 | 0.4625 | 0.4749 | 0.5375 | 0.5251 |
|                | Bagging          | 4 | 0.5400 | 0.5663 | 0.4600 | 0.4337 |
Table 4. Performance comparison of feature selection methods based on soil and environmental conditions.

| Feature Method | Selection       | Classification Method | Attribute Selected | Accuracy | Error Rate |
|----------------|-----------------|-----------------------|--------------------|----------|------------|
|                |                 |                       | Before Reduction   | After Reduction | Before Reduction | After Reduction | Before Reduction | After Reduction |
| Boruta        | kNN             |                       | 0.6                | 0.6189   | 0.4        | 0.3811       |
| Naive Bayes   | 0.7854          | 0.7864                | 0.2146             | 0.2136   |
| Decision Tree | 0.8188          | 0.8193                | 0.1812             | 0.1807   |
| SVM            | 0.8312          | 0.8337                | 0.1688             | 0.1663   |
| Random Forest | 0.8875          | 0.8889                | 0.1125             | 0.1111   |
| Bagging       | 0.9062          | 0.91                  | 0.0938             | 0.09     |
| Sequential Forward Feature Selection | 16 | 13 | 0.6 | 0.6237 | 0.4 | 0.3763 |
| Naive Bayes   | 0.7854          | 0.788                 | 0.2146             | 0.212    |
| Decision Tree | 0.8188          | 0.8203                | 0.1812             | 0.1797   |
| SVM            | 0.8312          | 0.8471                | 0.1688             | 0.1529   |
| Random Forest | 0.8875          | 0.889                 | 0.1125             | 0.111    |
| Bagging       | 0.9062          | 0.9126                | 0.0938             | 0.0874   |
| Recursive Feature Elimination | 16 | 14 | 0.6 | 0.6321 | 0.4 | 0.3679 |
| Naive Bayes   | 0.7854          | 0.7917                | 0.2146             | 0.2083   |
| Decision Tree | 0.8188          | 0.8241                | 0.1812             | 0.1759   |
| SVM            | 0.8312          | 0.8572                | 0.1688             | 0.1428   |
| Random Forest | 0.8875          | 0.8974                | 0.1125             | 0.1026   |
| Bagging       | 0.9062          | 0.9272                | 0.0938             | 0.0728   |

also improve the prediction accuracy. Tables 2, Tables 3 and Tables 4 show that the RFE with bagging has the best crop prediction accuracy, based on both soil and environmental characteristics, compared to others based only on soil characteristics and environmental factors.

3.3.4 Performance evaluation of Feature selection method with the Bagging classifier using n-fold validation

From the results above, it is evident that all the feature selection techniques with the bagging classifier work better than other classifier methods with feature selection techniques. To validate the performance of the feature selection technique for crop prediction, two validation methods are used: fold variation and training-testing data splitting.

Table 5 shows a performance evaluation of feature selection methods with the bagging technique to find the most suitable crop for a particular land area. The representation is based on various cross-fold validation techniques to obtain the best fold of all wrapper feature selection methods. The ranges vary from 10 to 90 folds, and the performance is evaluated using the metrics of Accuracy, Kappa, Precision, Recall, Specificity, F1 Score, MAE, and Log Loss.

Table 5 shows that the performance of all feature selection techniques with the bagging classifier working best with the 10 fold variation. From the result, it is observed that, for all the performance metrics the RFE technique outperforms the other two techniques.

3.3.5 Performance evaluation of Feature selection method with the Bagging classifier using data splitting validation

Table 6 shows a performance evaluation of feature selection methods with bagging techniques to find the most suitable crop for a particular land area. The representation is based on data splitting to get the best training and testing splitting ranges. The ranges vary from 25% – 75% to 75% – 25%. The performance is evaluated using the metrics of Accuracy, Kappa, Precision, Recall, Specificity, F1 Score, MAE, and Log Loss.
### 4. Conclusion

Data mining is a scientific field with applications in the study of crop yields. The prediction of crop cultivation is critical to agriculture, with farmers keen to work out how much they can possibly expect to produce. In the past, cultivar prediction was carried out by taking into account farmers’ basic understanding of specific stretches of land and the crops to be grown therein. In agriculture, several data mining strategies are used and analysed to predict crop cultivation in the future. This research has focused on comparing the feature selection methods of different
Table 6. Performance comparisons of feature selection methods with the bagging classifier based on data splitting validation.

| Metrics     | Performance of feature selection methods based on data splitting range (%) |          |          |          |
|-------------|---------------------------------------------------------------------------|----------|----------|----------|
|             |                                                                           | RFE      | Boruta   | SFFS     |
| Accuracy    |                                                                           |          |          |          |
| 25–75       | 73.58                                                                     | 72.84    | 72.92    |          |
| 30–70       | 76.27                                                                     | 76.07    | 76.13    |          |
| 35–65       | 81.43                                                                     | 80       | 80.29    |          |
| 40–60       | 86.11                                                                     | 84.74    | 85.98    |          |
| 45–55       | 89.41                                                                     | 88.78    | 88.44    |          |
| 50–50       | 88.93                                                                     | 86.35    | 86.72    |          |
| 55–45       | 91.7                                                                      | 90.7     | 90.95    |          |
| 60–40       | 90.65                                                                     | 89.52    | 89.79    |          |
| 65–35       | 91.89                                                                     | 90.83    | 91.03    |          |
| **70–30**   | **92.72**                                                                 | **91**   | **91.26**|          |
| 75–25       | 92.12                                                                     | 90.49    | 90.94    |          |
| Kappa       |                                                                           |          |          |          |
| 25–75       | 68.93                                                                     | 68.04    | 68.21    |          |
| 30–70       | 73.1                                                                      | 71.03    | 72.41    |          |
| 35–65       | 78.19                                                                     | 76.51    | 76.85    |          |
| 40–60       | 84.34                                                                     | 82.07    | 83.55    |          |
| 45–55       | 87.22                                                                     | 86.83    | 86.43    |          |
| 50–50       | 87.04                                                                     | 84.02    | 84.44    |          |
| 55–45       | 90.27                                                                     | 89.27    | 89.73    |          |
| 60–40       | 89.05                                                                     | 88.24    | 88.14    |          |
| 65–35       | 90.49                                                                     | 89.13    | 90       |          |
| **70–30**   | **91.94**                                                                 | **90.22**| **90.73**|          |
| 75–25       | 91.08                                                                     | 89.72    | 89.92    |          |
| Precision   |                                                                           |          |          |          |
| 25–75       | 80.22                                                                     | 80.07    | 80.11    |          |
| 30–70       | 82.92                                                                     | 82.73    | 82.82    |          |
| 35–65       | 86.94                                                                     | 86.28    | 86.46    |          |
| 40–60       | 88.72                                                                     | 87.37    | 88.01    |          |
| 45–55       | 90.25                                                                     | 89.29    | 89.91    |          |
| 50–50       | 91.45                                                                     | 89.38    | 89.94    |          |
| 55–45       | 94.73                                                                     | 93.98    | 94       |          |
| 60–40       | 92.56                                                                     | 91.87    | 92.09    |          |
| 65–35       | 93.88                                                                     | 92.53    | 93.11    |          |
| **70–30**   | **95.13**                                                                 | **94.32**| **94.99**|          |
| 75–25       | 94.74                                                                     | 94.15    | 94.21    |          |
| Recall      |                                                                           |          |          |          |
| 25–75       | 78.84                                                                     | 78.15    | 78.73    |          |
| 30–70       | 81.45                                                                     | 81.09    | 81.36    |          |
| 35–65       | 85.39                                                                     | 84.31    | 84.61    |          |
| 40–60       | 86.18                                                                     | 85.09    | 86       |          |
| 45–55       | 89.74                                                                     | 88.97    | 89.16    |          |
| 50–50       | 90.86                                                                     | 89.1     | 89.21    |          |
| 55–45       | 93.67                                                                     | 92.35    | 92.98    |          |
| 60–40       | 92.41                                                                     | 91.04    | 91.56    |          |
| 65–35       | 93.57                                                                     | 92.12    | 92.89    |          |
| **70–30**   | **94.34**                                                                 | **93.33**| **93.88**|          |
| 75–25       | 94.18                                                                     | 93.03    | 93.54    |          |
| Specificity |                                                                           |          |          |          |
| 25–75       | 96.51                                                                     | 96.41    | 96.49    |          |
| 30–70       | 96.85                                                                     | 96.69    | 96.78    |          |
| 35–65       | 97.54                                                                     | 97.35    | 97.38    |          |
| 40–60       | 97.91                                                                     | 97.79    | 97.85    |          |
| 45–55       | 98.35                                                                     | 98.13    | 98.29    |          |
| 50–50       | 98.55                                                                     | 98.22    | 98.26    |          |
| 55–45       | 98.91                                                                     | 98.8     | 98.85    |          |
| 60–40       | 98.78                                                                     | 98.67    | 98.7     |          |
| 65–35       | 98.94                                                                     | 98.83    | 98.88    |          |
| **70–30**   | **99.11**                                                                 | **99.01**| **99.09**|          |
| 75–25       | 99.09                                                                     | 98.89    | 98.94    |          |

(Continued)
prediction models, and suggested the best method to forecast crop cultivation in the future. Our findings from the results obtained conclusively show that Boruta, SFFS, and RFE feature selection techniques with the bagging classifier performing best with 10 fold and 70% – 30% data splitting range and RFE with the bagging method, outperforms other methods.

**Acknowledgments**

We would like to thank the Department of Agriculture Sankarankovil Taluk, Tenkasi District, Tamilnadu, India for providing data for the analysis.

**Disclosure statement**

No potential conflict of interest was reported by the authors.
References

[1] S. Das. Filters, wrappers and a boosting-based hybrid for feature selection. In International conference on machine learning 1 (2001), pp. 74–81.

[2] H. Liu and Y. Lei, Toward integrating feature selection algorithms for classification and clustering, IEEE Trans. Knowl. Data Eng. 17 (2005), pp. 491–502. doi:10.1109/TKDE.2005.66.

[3] A. Maruf, M. Abdullah, and S. Shatabda, iRSpot-SF: Prediction of recombination hotspots by incorporating sequence based features into Chou’s Pseudo components, Genomics 111 (2019), pp. 966–972. doi:10.1016/j.ygeno.2018.06.003.

[4] I. Kononenko, Estimating attributes: Analysis and extensions of RELIEF. European conference on machine learning 784 (1994), pp. 171–182.

[5] J. Pavel Pudil, N.N. Choakjarernwanit, and J. Kittler, Feature selection based on the approximation of class densities by finite mixtures of special type, Pattern Recognit 28 (1995), pp. 1389–1398. doi:10.1016/0031-3203(94)0009-B.

[6] J. Novovicová, P. Pudil, and J. Kittler, Divergence based feature selection for multimodal class densities, IEEE Trans. Pattern Anal. Mach. Intell. 18 (1996), pp. 218–223. doi:10.1109/34.481557.

[7] H. Liu and H. Motoda, Feature Selection for Knowledge Discovery and Data Mining, 1st ed., Springer, US, 1998.

[8] M. Dash, K. Choi, P. Scheuermann, and H. Liu. Feature selection for clustering-a filter solution. IEEE International Conference on Data Mining, 2002, Proceedings., Maebashi City, Japan, pp. 115–122.

[9] I. Guyon, S. Gunn, M. Nikravesh, and L.A. Zadeh (eds.), Feature Extraction: Foundations and Applications, 1st ed., Springer-Verlag, Berlin Heidelberg, 2008.

[10] I. Guyon and A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003), pp. 1157–1182.

[11] W. Paja, K. Pancerz, and P. Grochowalski, Generational feature elimination and some other ranking feature selection methods, in Advances in Feature Selection for Data and Pattern Recognition, International Publishing, Springer,Charm, Vol. 138, 2018, pp. 97–112.

[12] J.-Y. Hsieh, W. Huang, H.-T. Yang, C.-C. Lin, Y.-C. Fan, and H. Chen. Building the rice blast disease prediction model based on machine learning and neural networks., Easy chair the world of scientists (2019), pp. 1–8.

[13] R. Kohavi and G.H. John, Wrappers for feature subset selection, Artif. Intell. 97 (1997), pp. 273–324. doi:10.1016/S0004-3702(97)00043-X.

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, Gene selection for cancer classification using support vector machines, Mach Learn 46 (2002), pp. 389–422. doi:10.1023/A:1012487302797.

[15] M. Sebban and R. Nock, A hybrid filter/wrapper approach of feature selection using information theory, Pattern Recognit 35 (2002), pp. 835–846. doi:10.1016/S0031-3203(01)00084-X.

[16] Y. Lei and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In Proceedings of the 20th International Conference on Machine Learning, Washington, DC, USA, 2003.

[17] P. Somol, J. Novovicová, and P. Pudil. Flexible-hybrid sequential floating search in statistical feature selection, in Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR), Dy. Yeung, J. T. Kwok, A. Fred, F. Roli, D. de Ridder eds., Vol. 4109, Springer, Berlin, Heidelberg, 2006, pp. 632–639.
[18] A. Salappa, M. Doumpos, and C. Zopounidis, Feature selection algorithms in classification problems: An experimental evaluation, Optimisation Methods and Software 22 (2007), pp. 199–212. doi:10.1080/10556780600881910.

[19] M.B. Kursa, A. Jankowski, and W.R. Rudnicki, Boruta—a system for feature selection, Fundamenta Informaticae 101 (2010), pp. 271–285. doi:10.3233/FI-2010-288.

[20] A. Marcano-Cedeno, J. Quintanilla-Dominguez, M.G. Cortina-Januchs, and D. Andina, Feature selection using sequential forward selection and classification applying artificial metaplasticity neural network. In Industrial electronic conference, 36th annual conference on IEEE industrial electronics society, Glendale-USA, 2010.

[21] Z. Karimi, M.M.R. Kashani, and A. Harounabadi, Feature ranking in intrusion detection dataset using combination of filtering methods, International Journal of Computer Applications 78 (2013), pp. 21–27. doi:10.5120/ijca2016909966.

[22] S. Chouhan, D. Singh, and A. Singh, An improved feature selection and classification using decision tree for crop datasets, International Journal of Computer Applications 142 (2016), pp. 5–8. doi:10.5120/ijca2016909966.

[23] D. Heckmann, U. Schlüter, and A.P.M. Weber, Machine learning techniques for predicting crop photosynthetic capacity from leaf reflectance spectra, Mol Plant 10 (2017), pp. 878–890. doi:10.1016/j.molp.2017.04.009.

[24] P.S. Maya Gopal and R. Bhargavi., Feature selection for yield prediction in Boruta algorithm, International Journal of Pure and Applied Mathematics 118 (2018), pp. 139–144.

[25] A. Bahl, B. Hellack, M. Balas, A. Dinischiotu, M. Wiemanne, J. Brinkmann, A. Luch, B. Y. Renard, and A. Haase, Recursive feature elimination in random forest classification supports nanomaterial grouping, Nanolmpact 15 (2019), pp. 100179. doi:10.1016/j.immpact.2019.100179.

[26] P.S. Maya Gopal and R. Bhargavi., Performance evaluation of best feature subsets for crop yield prediction using machine learning algorithms, Appl. Artif. Intell. 33 (2019), pp. 621–642. doi:10.1080/08839514.2019.1592343.

[27] P.S. Maya Gopal and R. Bhargavi., Selection of important features for optimizing crop yield prediction, International Journal of Agricultural and Environmental Information Systems (IJAEIS) 10 (2019), pp. 54–71. doi:10.4018/IJAEIS.2019070104.

[28] G.B. Dela Cruz, B.D. Gerardo, and B.T. Tanguilig, Agricultural crops classification models based on PCA-GA implementation in data mining, International Journal of Modeling and Optimization. 4 III, (2014), pp. 375–382. doi:10.7763/IJMO.2014.V4.404.

[29] V.N. Vapnik, Statistical Learning Theory Wiley, New York, 1998.

[30] D.H. Zala and M.B. Chaudhri, Review on use of BAGGING technique in agriculture crop yield prediction, International Journal for Scientific Research & Development 6 (2018), pp. 675–677.

[31] J. Camargo and A. Young, Feature selection and non-linear classifiers: Effects on simultaneous motion recognition in upper limb, IEEE Trans. Neural. Syst. Rehabil. Eng. 27 (2019), pp. 743–750. doi:10.1109/TNSRE.2019.2903986.

[32] F. Shirbani and H. Soltanian Zadeh, Fast SFFS-based algorithm for feature selection in biomedical datasets, Amirkabir International Journal of Science& Research (Electrical & Electronics Engineering) 45 (2013), pp. 43–56.

[33] P.M. Granitto, C. Furlanello, F. Biasioli, and F. Gasperi, Recursive feature elimination with random forest for PTR-MS analysis of agroindustrial products, Chemometrics and Intelligent Laboratory Systems 83 (2006), pp. 83–90. doi:10.1016/j.chemolab.2006.01.007.

[34] F. Balducci, D. Impedovo, and G. Pirlo, Machine learning applications on agricultural datasets for smart farm enhancement, Machines 6 (2018), pp. 38–59. doi:10.3390/machines6030038.

[35] R. Jahan, Applying Naive Bayes classification technique for classification of improved agricultural land soils, International Journal for Research in Applied Science & Engineering Technology (IJRASET) 6 (2018), pp. 189–193. doi:10.22214/ijraset.2018.5030.

[36] J.A.K. Suykens and J. Vandewalle, Least squares support vector machine classifiers, Neural Processing Letters 9 (1999), pp. 293–300. doi:10.1023/A:1018628609742.
[37] D. Anantha Reddy, B. Dadore, and A. Watekar, *Crop recommendation system to maximize crop yield in Ramtek region using machine learning*, International Journal of Scientific Research in Science and Technology 6 (2019), pp. 485–489. doi:10.32628/IJSRST196172.

[38] M. Lango and J. Stefanowski, *Multi-class and feature selection extensions of roughly balanced bagging for imbalanced data*, J Intell Inf Syst 50 (2018), pp. 97–127. doi:10.1007/s10844-017-0446-7.