RESEARCH ARTICLE

INTERNATIONAL RESEARCH JOURNAL OF MULTIDISCIPLINARY TECHNOVATION

Asian Research Association

DOI: 10.54392/irjmt2525

Empowering Precise Crop Recommendation System by Accompanying Tree Covariance Matrix-Parallel Random Forest Classifier

Umamaheswari R [a, *], E. Kannan [a]

[a] Department of Computer Science and Engineering, Vel Tech Rangarajan Dr.Sagunthala R & D Institute of science and Technology, Chennai, Tamil Nadu, India.

* Corresponding Author Email: uma2007ap83@gmail.com

DOI: https://doi.org/10.54392/irjmt2525

Received: 10-01-2025; Revised: 26-02-2025; Accepted: 07-03-2025; Published: 17-03-2025

Abstract: Transformation in crop management systems, particularly in creating an environment that gives rise to sustainable farming, is achieved due to innovation and the advancement of modernized agricultural technology. Anyhow, meeting the increasing food demand is one of the great challenges that stand in front of the farmers. By taking into account, factors like soil, climate, and seasonality, the crop recommendation system plays a central role in providing customized guidance to the farmers. Current crop recommendation models are often confined by a paucity of feature selection, spatial-temporal integration shortfalls, and a finite amount of decision-tree diversity. All these shortfalls retrain their scalability and accuracy. To overcome the aforementioned blocks, an innovative framework is projected that includes the Best Incremental Random Subset (BIRS) feature selection method for choosing the best features and the Parallel Random Forest (PRF) -Tree Covariance Matrix model (PRF-TCM) encourages decision-tree diversity, permitting more accurate and efficient crop recommendations. Experimental results reveal that the proposed framework outperforms existing models with accuracy (89.7), precision (88.6), and recall (87.5). The framework shows significant improvements over current models, responsible for more viable agricultural practices.

Keywords: BIRS, Crop Recommendation, Machine Learning, Parallel Random Forest, Tree Covariance Matrix

1. Introduction

The mainstay of human civilization is agriculture, which plays a significant role in revitalizing and sustaining global food systems [1]. As a result of speedy changes in climate, soil health, and universal agricultural demand, the urgency in acquiring data-driven solutions for crop recommendation has been more [2, 3]. Leading technologies such as machine learning, artificial intelligence, and big data are employed in crop a recommendation system that embellishes agricultural productivity and sustainability [4, 5]. By utilizing various factors like soil type, climatic conditions, and seasonality, these techniques recommend an optimal crop. Specifically, the upswing in precision agriculture has brought these techniques as a cutting-edge technology in agricultural inventions [6, 7].

The current crop recommendation system, entrust on very simplified approaches that disregard the critical relationship between soil health, climate, and crop compatibility [8, 9] Additionally, many models stuck to look at the diversity and variability in agricultural datasets, which leads to flawed crop suggestions [10]. Moreover, choosing the relevant features from a huge dataset remains a difficult and unsettled issue that leads to over fitting and poor generalization [11]. So prosperous; recommendation system is needed to blend a collection of datasets to contribute to diverse crop recommendations [12, 13].

To tackle crop recommendation problems, many existing algorithms, like Random Forest (RF) [14-16], Support Vector Machines (SVM), Decision Tree (DT) [17-19], Gaussian Naïve Bayes (GNB) and Neural Networks are used to anticipate suitable crops. Other than these methods, clustering algorithms are also used in the recommendation process that segregate clusters of crops based on akin characteristics, surroundings and so on [20] Regardless of the evolution done by the existing algorithms, they still suffer from several flaws [21, 22]. One of the major shortfalls is failure to address the fundamental variability in soil and climatic factors, which influence crop prediction accuracy. One more dare is the absence of diversification among decision trees in models, that consequences robustness of the recommendations. Moreover, many models fail to offer a comprehensive solution that integrates both spatial

and temporal factors, which are critical for making accurate crop predictions across different seasons and regions.

To suppress these restraints, an innovative framework PRF-TCM is used to enhance the accuracy and robustness of crop recommendation systems and Best Incremental Random Subset (BIRS) for feature selection is proposed. In contrast with existing methods, that heavily depends on feature-based DT, proposed model incorporate covariance-driven mechanism to improve tree diversity, reduce redundancy and optimize feature interactions. By assigning adaptive weights to DT based on pairwise covariance, ensure to capture diverse agronomic patterns. By leveraging various soil and climate feature subsets, model regularly balances computational efficiency and predictive performance. Skillfully, the PRF takes advantage of handling huge datasets by utilizing multiple decision trees. The TCM assures inter-tree diversity and improved generalization. Along with this, the BIRS feature selection technique guarantees that only the required features are considered, which diminishes noise and enhances model predictive power. More over experimental results also narrates that model outperforms existing models in terms of accuracy, precision and recall.

Major contributions of the proposed work can be summarized as follows:

- A bountainous understanding of complex relationships between different agricultural factors is achieved by PRF-TCM.
- BIRS-based feature selection method guarantees that only the most relevant features are considered, to embellish prediction accuracy and model interpretability.

The research article is organized as follows: Chapter 1 provides the introduction, outlining the background, problem statement, and objectives of the proposed work. Chapter 2 reviews existing literature and highlights gaps that our research aims to fill. Chapter 3 explains the proposed methodologies in detail, focusing on the innovative aspects of the framework. Chapter 4 discusses the results and evaluates the performance of the proposed system through various metrics and compares it with existing approaches. Chapter 5 conclusion and future directions for improving the crop recommendation system.

2. Related Works

Kiran et al [17] uses data such as weather, soil pH, nutrient levels (Nitrogen (N), Potassium (P), and Phosphorous (K)), temperature, and rainfall to develop a machine learning-based crop prediction system with the support of algorithms like Gaussian Naïve Bayes (GNB), Support Vector Machine (SVM), Random Forest (RF), and Decision Tree (DT). This methodology improves productivity and efficiency. Though GNB classifier obtained high accuracy, this method does not fully account for socio-economic factors, influencing farming decisions. Crop cultivation is influenced by soil nutrition deficiency, climatic change and many more, so Maheswary *et al* [18] designed a crop recommender system using entropy and Gini index as performance metrics and validated using K-nearest neighbor (KNN), DT and RF classifiers. RF outperformed well than rest of the models Work addresses critical agricultural challenges with data-driven insights. Shortfall found here is, failed to address potential integration with real-time systems.

Apat *et al* [23] developed an AI-based decision support system for the soil nutrition and climate dataset. For data balancing, Synthetic Minority Over-Sampling Technique (SMOTE) was applied, and an optimization technique is used to enhance model performance. Cat Boost algorithm (C-Boost) achieved the best accuracy and F1-score of 0.9916. The focal point of this work is the use of AI and machine learning for crop recommendation. But this work heavily depends on dataset quality, which may not suit for all conditions. Mahmoud *et al* [24] designed an intelligent eXplainable Artificial Intelligence (XAI)–CROP system, which assists farmers in crop selection. This method surpasses all traditional algorithms with low Mean Squared Error (MSE).Also it beats other ML models in error metrics and interpretability. Restricted evaluation of real- world adaptability and farmer accessibility.

Dey *et al* [19] to recommend crops assess the efficacy of SVM, XGBoost, RF, KNN, and DT using soil nutrients and climatic factors. This approach suggests a potential for artificial intelligence (AI)-based interfaces to support speedy crop and fertilizer recommendations under fluctuating environmental conditions. Method separates agricultural and horticultural datasets for better prediction. But limited real-world confirmation and accessibility for farmers without digital tools. With the support of a high-resolution dataset, Burdett *et al* [14] studied and evaluated the performance of Multiple Linear Regression (MLR), Artificial Neural Network (ANN), DT and RF to predict corn and soybean yield. While MLR performed worst, cross-validation shows that random forest predict yield variability of untrained fields under specific conditions. Significant soil attributes and topographic features were identified, though predictions had uncertainties. Recognize key soil and topographic attributes influencing yield. Dependence on similar soil-yield relationships for untrained fields reduces generalizability.

Using machine learning, global positioning system (GPS), and cloud-based data storage, senapaty *et al* [25] introduced a personalized decision support system for crop recommendation. An Android app was developed to predict crops based on specific local regions. Farmers get fast and accurate crop

recommendations for their localized farms. The chances of getting a false positive rate might be more as the rate of specificity is very low. Rani *et al* [26] designed a hybrid algorithm, Long Short-Term Memory (LSTM) Recurrent Neural Networks (RNN) for predicting weather conditions and RF is used for crop selection. RF achieves high accuracy and fast processing time. Extensive crop selection is achieved, as it considers weather and soil data. Other metrics related to weather prediction include expecting RMSE.

Reddy and Kumar [3] use Geographically Weighted Random Forest Regression (GWRFR) to predict corn yield. Modal effectively addresses spatial non-stationary, indicated by lower Moran's I value in its residuals. But it focuses on specific country predictions. An intelligent crop monitoring is designed by yuan *et al* [11] which connects feature expansion and feature importance analysis employed to optimize the raw data. The model can predict classifications without all time-series data, enabling early monitoring. LSTM networks are complicated and computationally costly to train, requiring significant computational resources.

Janrao *et al* [16] introduces conglomerate crop recommendations, which suggest a cluster of crops instead of a single crop. Ensemble supervised clustering techniques (ESCT) algorithm comprises of K means algorithm and an inter-cluster correlation coefficient (ICCC).K values are optimized with support of an approximation function. The model learns through Back propagation algorithm. The use of function approximation to optimize the k- values improves convergence. But computational complexity is more. Cascade parallel random forest (CPRF) was used by Zhang et al [12] to evaluate rice diseases. Generally, the agriculture dataset contains imbalanced and high-dimensional data. To lighten this issue, CPRF is used. Make use of the spark platform for efficient processing of bulk datasets. Multiple random forests likely boost computational complexity compared to a single random forest.

To recommend something suitable for the regions where crop yield is low Saritha *et al* [27] use RF and hyper parameter tuning techniques. Popularize precision farming in low-yield areas. Interpretation depends on a single dataset, which may not fully represent the diversity of real-world farming conditions. Selecting the suitable features is an important step, so Gupta *et al* [8] takes in hand, Relief feature selection technique to pick the best features and extract features using Linear Discriminant Analysis (LDA). SVM, KNN, RF are engaged in order to predict the appropriate crop based on location, season and market demand. As it suggests, crop based on the current market situation, both farmers and policymakers are benefited. But comparison of classifiers with existing algorithm is not done. Downfall in agriculture arises due to improper selection of crop for cultivation, reddy *et al* [3] taken in hand collaborative filtering techniques to recommend crops based on soil and weather conditions. Statistical analysis and predictive modeling are applied in order to predict a suitable crop. Traditional models, such as RF [17-19, 14-16]. DT [17-19, 14] utilize heuristic-based feature selection, leading to repetition and shrinkage efficiency of the model. These models often struggle to handle high-dimensional agricultural datasets. Hybrid models like GWRFR [11], CPRF [12] lack systematic feature elimination which affects interpretability and robustness.

Table 1. Comparison of existing algorithms strength and weakness

Algorithms	Strength	Weakness
SVM	Handles high-dimensional dataset	Difficult to handle imbalanced datasets
RF	Robust and interpretable	N number of trees leads to over fitting
XGBoost	High accuracy and much powerful in training the data.	Computationally much expensive
CNN	High feature learning capability	Needs large dataset.
GNB	Computation is faster, works well even with small dataset, effective for normally distributed features	Assumes feature independence, limited performance on complex datasets
C-Boost	Efficient in handling categorical features, prevents over fitting, robust on unbalanced data	High in computational cost
GWRFR	Catches spatial dependencies, flexible to geographical variations	Computationally costlier, sensitive to feature scaling
PRF-TCM	Enhance model diversity, robust crop prediction, reduces over fitting	Requires hyper parameter Tuning, sensitive to feature correlation

Relief feature selection [8] follows static feature selection, so it fails to adapt dynamically for various datasets. So that feature selection techniques like BIRS are necessary to improve scalability and interpretability. Though SMOTE can handle imbalanced data, but it may produce synthetic data that do not fully denotes real-world conditions. Even though ensemble models like XGBOOST [19] strengthen robustness, but scarcity in diversity among decision trees reduces their generalization ability. So to improvise the diversity of trees in PRF, TCM construction is done. The research gap identified can be overcome by developing a robust, scalable, and interpretable crop recommendation framework. The motivation behind this work is to develop a novel framework to address the identified gaps. The proposed framework employs BIRS feature selection method to dynamically select relevant features, to improve scalability, accuracy and enable large-scale crop recommendations. PRF-TCM ensures diversity between various tress which lessen over fitting issues of RF and XGBoost. Table 1 shows the strength and weakness of existing and proposed algorithms.

3. Proposed Model

Accompanying PRF-TCM is proposed in this work. As discussed in the related work, many existing methodologies are there to recommend crop, but still to improve accuracy and ensure diversity this model is developed.

3.1 Data Preprocessing

Preprocessing is the first and critical step, which ensures high-quality input data before training machine learning models. In the PRF-TCM framework, below mentioned preprocessing steps are carried out:

Before the data are handled by machine learning algorithms, unprocessed data undergoes cleaning procedure to handle inconsistencies and errors. Outliers are identified using Z-score and the outliers are put back with median values.

$$z = (x - \mu)/\sigma \quad (1)$$

In equation (1) z indicates z-score values, x is the value of element, μ denotes mean population.,σ is standard deviation.

Missing values weaken performance of machine learning models as it leads to bias. To identify the missing values, features with missing values are determined using Pearson correlation with other variables to determine imputation feasibility.

$$r = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2 \ \sum(y-\bar{y})^2}} \quad (2)$$

Here (equation 2) x, y denotes the values of the variables, $\bar{x}, \bar{y}$ indicates mean value and square root represents the sum of the values.

To identify the missing values in categorical Features (Kharif, Rabi, Zaid), Mode imputation is used, in which, to fill the missing values most frequent category is used. Once data cleaning, is over, to remove irrelevant and redundant features and to identify the most suitable features, BIRS feature selection techniques is used, that rank features based on importance scores.

3.2 PRF-TCM Model Architecture

Figure 1 portrays the architecture diagram of the proposed model. The model begins by BIRS feature selection technique to guarantee that appropriate features are included in the prediction process. Multiple decision trees are trained using subsets of features selected in PRF phase.

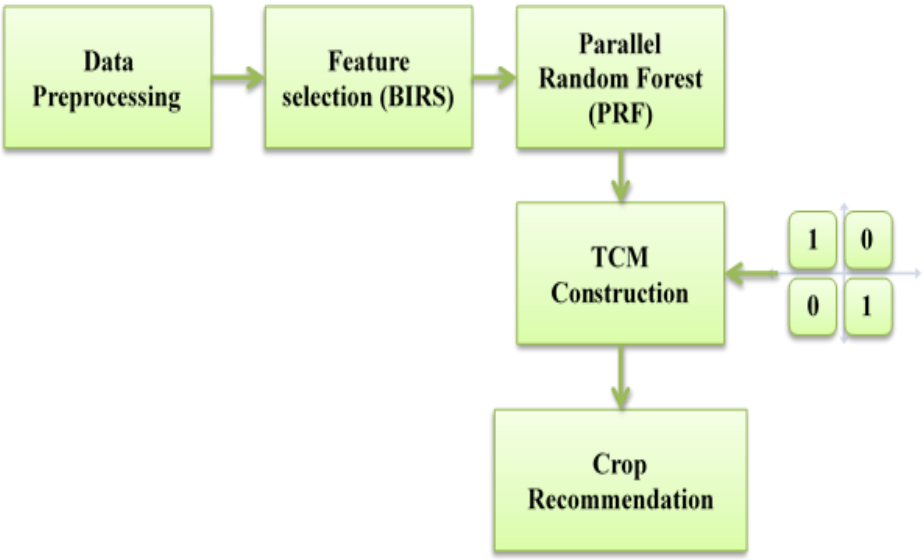

Figure 1. PRF-TCM Architecture

This approach grants the model to gain knowledge about various aspects of the data. To improve the performance of PRF, TCM is enforced to measure pairwise correlation between trees. The covariance matrix helps in assigning optimal weights to each tree based on its diversity, which is crucial for achieving better prediction accuracy.

3.3 Tree Covariance Matrix Based Parallel Random Forest for Crop Recommendation System

As aforesaid in the architecture diagram description, the proposed work aims to enhance crop recommendation with BIRS, PRF-TCM. In particular, the framework combines the strengths of feature selection using BIRS and to predict the most appropriate crop PRF–TCM methodology is used. In which, TCM is used to promote diversity among decision trees in the forest, improving prediction accuracy while addressing the high dimensionality of input data and PRF consists of five different , each RF, works with one particular dataset that are discussed in the result and discussion section.

3.3.1 BIRS Feature Selection

The most appropriate features for the proposed model are consistently picked up by BIRS. Weights are assigned to features by computing mutual information and features are placed in decreasing order to estimate those with the highest impact on crop recommendation accuracy. Model performance in evaluated using accuracy, precision, recall, when new features are added. Processes stopped, when a new feature does not improve model performance, and repetitious features are discarded. The approach effectively handles high- dimensional data and boosts up the efficiency of the PRF. Algorithm 1 narrated down elaborates the working principle of BIRS.

Algorithm1. BIRS feature selection

Input

fn →Features in dataset (soil type, soil fertility, soil nutrition, climate factors, seasonality)

Cp →Target variable (crop)

Output

$f_{selection}$ →Selected features from f_n

Steps

1. Feature importance ($importance(fn)$) calculation
2. Compute mutual information mi
3. mi=Entropy (*fn*) + Entropy (Cp) - Entropy (*fn*). Entropy (Cp)
4. Assign weight (w) to fn based on their $importance(fn)$
5. $importance(fn)$ = mi
6. Rank features descending order of their w
7. Initialize ($f_{selection}$← ( ),best_performance ←0, current_ fn ←( ))
8. Do iterative $f_{selection}$
9. for each feature fn in the ranked list:
10. Add fn ←current_ fn
11. Train model ←current_ fn and evaluate performance ( )
12. if evaluate performance ( ) > threshold
13. Update best_performance
14. Add fn← $f_{selection}$
15. else
16. Remove fn from current_ fn
17. Remove redundant features
18. Return $f_{selection}$

Selection of features is done by BIRS feature selection technique, which is formulated as:

$$f_{selection} = max\{f1,2..fn\}\, i = 1^{importance(fn)} \quad (3)$$

f selection indicates the features that are selected and *importance(fn)* relevance score of

f selection, which is computed based on its ability to improve model performance. Equation (3) ensures that only the most important features are included in the model, thereby reducing computational complexity.

3.3.2 Working Principle of Prf-Tcm

Each decision tree DT_i in the PRF model is trained on a subset of features $f_{selection,}$

$$DT_i = (fselection, Cp) \quad (4)$$

Where $f_{selection}$ represents the subset of features for tree and C_p is the target crop type. Equation (4) highlights how trees in the PRF are trained independently on different feature subsets, allowing the model to learn specialized representations of the data. For each tree DT_i in the PRF, generate predictions $DT_i(Xk)$ for a validation dataset Multiple decision trees generate multiple outputs. So prediction output of the trees, DT_i and DT_j for the validation dataset is computed.

$$.DT_i(X) = [T_{i1},\ T_{i2}\ , \ldots T_{iN}] \quad (5)$$

$$DT_j(X) = [\,T_{j1}, T_{j2}\ldots . T_{jN}] \quad (6)$$

Where *N* in total number of samples in *X*. The covariance between predictions from trees i and j is given by:

$$CovM(DT_i, DT_j) = \frac{((DT_i(X_k) - \overline{DT_i})(DT_j(X_k) - \overline{DT_j}))}{m-1} \quad (7)$$

In which $DT_i(Xk)$ and $DT_j(Xk)$ are predictions from trees *i* and *j* for sample X_k and

$\overline{DT_i}$ *and* $\overline{DT_i}$ are average predictions of trees *i* and *j* over the entire validation set. Equation (7) defines how to calculate the covariance between pairs of trees, which is used to construct the symmetric Tree Covariance Matrix. Pearson correlation (PCorrelation) is computed based on,

$$PCorrelation\left(DT_i(X), DT_j(X)\right) = \frac{CovM(DT_i(X), DT_j(X))}{\sigma(DT_i(X), DT_j(X))} \quad (8)$$

In equation (8), $CovM(DT_i(X), DT_j(X))$ is the covariance between the predictions, $\sigma(DT_i(X))$ and $\sigma(DT_j(X))$ is the standard deviation of $(DT_i(X))$ and $DT_j(X)$ respectively.

Algorithm 2 Tree Covariance Matrix (TCM) construction for PRF

Input:

Dataset with features *fn* and Target variable *Cp*

Training and validation subsets Number of trees T

Output:

Weighted crop predictions using TCM based PRF.

1. Train PRF
2. for (i= 1 to T)
3. Randomly pick feature from *fselection*
4. Train a *DTi* using *fselection*
5. for each, *DTi* predict crop and load results in *DTi* , *DTj*
6. To construct TCM, Initialize a N×N matrix with zero
7. for i=1 to T:
8. for j=1 to T:
9. Compute← $PCorrelation\ (DT_i(X), DT_j(X))$
10. set $T_{cm(i,j)}$= $PCorrelation\ (DT_i(X), DT_j(X))$
11. set diagonal entries of $T_{cm(i,j)}$ = 1
12. compute ←wt_i
13. Normalize wt_i
14. Apply weighted voting
15. Return (*Cp*)

The TCM acts as the stamina of the proposed model by sanctioning the PRF to utilize diverse and complementary decision trees energetically. Outcome of this is robust, accurate, and context-aware crop recommendations that beat traditional approaches confined to uniform tree weighting. To construct the TCM, design a matrix T_{cm} of size N* N times, where N is the sum of tree in PRF. Each element in Tcm(i,j) indicates the covariance between tree DT_i and DT_j.

$$T_{cm(i,j)} = \begin{pmatrix} Pcorrelation(DT_i(X), DT_j(X), i \neq j \\ 1 \quad , \quad i = j \end{pmatrix} \quad (9)$$

As $\boldsymbol{Pcorrelation\ (DT_i, DT_j)}$ = $\boldsymbol{Pcorrelation\ (DT_j, DT_i)}$ is equal, the TCM matrix is symmetric. For an instance, suppose if three trees (DT1, DT2, DT3) are there, then matrix representation will be like this. Based on the number of trees the matrix size will differ.

$$\begin{pmatrix} 1 & T_{cm12} & T_{cm13} \\ T_{cm21} & 1 & T_{cm23} \\ T_{cm31} & T_{cm32} & 1 \end{pmatrix}$$

Sum of covariance of each tree is find, by aggregating the covariance of each row in the matrix.

$$sum_i = T_{cm(i.j), i \in \{1,2 \ldots N\}} \quad (10)$$

Based on covariance value, weight (wti) of the tree is assigned. Therefore wti is assigned to each tree inversely proportional to their total covariance in equation (11).

$$wt_i = \frac{1}{sum_i}\ , i \in \{1,2, \ldots N\} \quad (11)$$

After an inverse value is computed, finally wti for each tree is computed based on the inverse of the sum of covariance values across all trees:

$$wt_i = \frac{1}{i-1^N CovM(DT_i, DT_j)} \quad (12)$$

In equation (12) N is the total number of trees. Assign higher weights to the trees that are less correlated with each other, promote diversity and improve prediction robustness. Normalization of the weights is done to bring their sum to 1. Finally crop recommendation is done.

4. Results and Discussion

In this section, narration about the dataset is debated, followed by this performance metrics are discussed. Lastly, the evaluation of the proposed model with existing algorithms are analyzed.

4.1 Summary of Dataset

To carry out the work, crop dataset is obtained from the Kaggle website [7]. Four different types of dataset, namely soil fertility, soil type, soil nutrition, climate data, and seasonal dataset are used to recommend crops. In each of the dataset, the target variable is the crop. Collectively all five dataset contains 30 features, but, from 30 features only 15 features: Nitrogen (N), Potassium (P), Phosphorous (K), temperature, humidity, rainfall, pH, EC, OC, OM, sand, silt, clay, CEC, CaCO3 are selected as best by the BIRS used throughout the work.

4.2 Performance Metrics

Performance metrics are used to assess the predictive talent of the developed model.

Introductions to the evaluation metrics are discussed in Table 2.

Table 2 represents purpose (usage) of each of the performance metrics accuracy, precision, recall. Purpose of these metrics is to measure efficiency of the model. Efficiency of the model is computed using the formula mentioned in the table 2.

4.3 Evaluation of Proposed Model

Each tree is trained with an exclusive subset of features and correlation between trees is predicted using TCM, which is helpful in evading similar predictions. On probation, the number of trees is raised until TCM denotes a plateau in added diversity, ensuring computational efficiency. Since deeper trees lead to overfitting and less deep trees might fail to identify some important relations, TCM is useful in finding an optimal tree depth. Table [3] measures the key parameters of the proposed framework with three existing algorithms, like SVM, RF, and GB.Interdependencies between features are effectively captured using BIRS.Its effectiveness is compared with existing features selection methods like PCA [28], Random search [29, 30]. BIRS based feature selection procedure outperforms well in all metric measures: accuracy (89.7), precision (88.6), recall (87.5).

Table 4, shows the comparison analysis of variations in feature selection techniques. In proposed model, 15 excellent features are selected to ensure balance between model simplicity and predictive power.

Selection of features is the main part in designing a model, as it improvises the model's performance. Table 5 demonstrates the accuracy remains steadily high, when the number of feature counts is gradually reduced.

Figure 2 illustrate the performance rate of the proposed model when handling different set of features and the model achieves highest accuracy rate (89.7), precision (88.6), and recall (87.5), when handling 15 features. Series 1, 2, 3 indicates accuracy, precision and recall respectively.

Table 2. Performance metrics

Metrics	Formula	Purpose
Accuracy	Accuracy = [TP+TN] [TP+TN+FP+FN]	Measures how frequently the model correctly anticipate an outcome
Precision	Precision = [TP] [TP+FP]	Express the model's capacity to prevent false positives.
Recall	Recall=[TP] [TP+FN]	Model’s ability to correctly predict the target class

Table 3. Comparison Table for Key Parameters

Algorithm	Feature Selection Method	Number of features	Number of DT	Tree Depth	Training - Testing Split (%)	Accuracy (%)	Precision (%)	Recall (%)
PRF - TCM	BIRS	**15**	**100**	**10**	**80-20**	**89.7**	**88.6**	**87.5**
SVM	Principle component analysis (PCA)	10	N/A	N/A	75-25	84.5	82	83.8
RF	Random Search	20	60	7	70-30	88.2	86.7	86.9
GB	Recursive Feature Elimination (RFE)	18	80	8	80-20	89.6	87	84.4

Table 5. Feature selection count variation performance analysis

Feature Selection Count	PRF-TCM	Accuracy (%)	Precision (%)	Recall (%)
25 Features	High correlation features retained	86.5	85.1	81.4
20 Features	Reduced irrelevant features	83.9	82.8	82.2
15 Features	Optimal selection	89.7	88.6	87.5
10 Features	Minimal features; slight performance drop	84.4	82.7	81

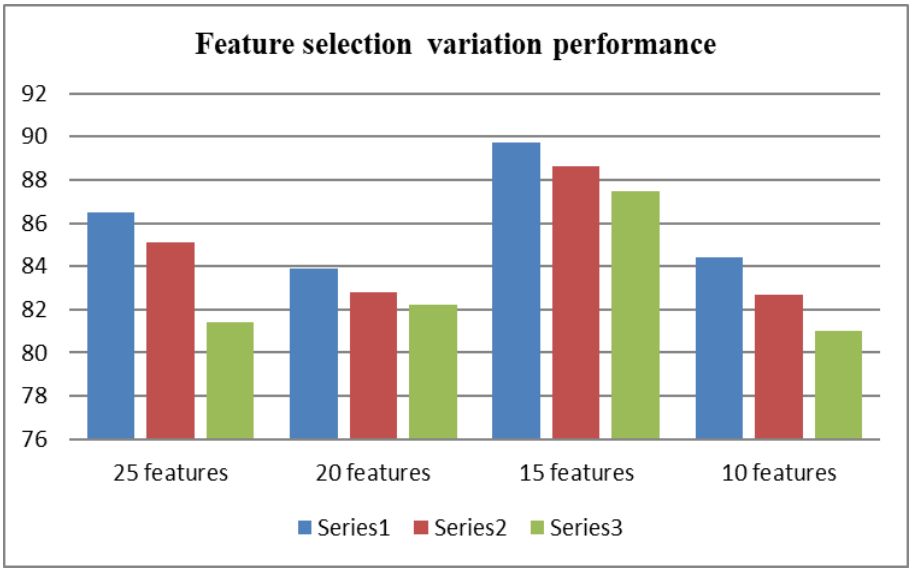

Figure 2. Feature selection variation performance over PRF-TCM

5. Conclusion

Work proposed in this paper combines PRF-TCM to boost the accuracy and efficiency of crop recommendation systems. Selection of the feature is influenced by the BIRS technique through which the feature selection process gets optimized. The PRF is trained with 15 selected features and each tree predicts output for the subset of features. Further to upgrade the performance of PRF, TCM is clubbed with PRF to manipulate correlation between the trees. Experimentally, the results prove that the proposed model outperforms existing algorithms in terms of accuracy (89.7), precision (88.6), and recall (87.5). When compared to other feature selection techniques, BIRS selects optimal features.

From the practical point of view, this framework exhibits a high degree of scalability and robustness, contributing to the advancement of sustainable agricultural practices and offering farmers a more reliable decision-making tool. Proposed model offers notable benefits to the farmers by providing data-driven insights for desirable crop selection, reduce confidence on instinct based decision making process. Farmers can gain high yield, as optimal crop is suggested based on soil fertility, soil nutrition, soil type climate conditions, and season. For policymakers, this framework enables the development of precision agriculture strategies, supports evidence-based policies for sustainable farming practices. .Integration of spatial and temporal factors, are critical for making accurate crop predictions across different seasons and regions. Future studies can blend spatial data like soil map with temporal data like season, yield to increase accuracy rate of crop recommendation. This blending procedure allow for dynamic crop recommendation according to changing environmental conditions. So in future work spatial – temporal factors can be added in the model to enhance accuracy and to suggest crop based on a localized area. Also integrating real-time sensor data and remote sensing technologies to future upgrade its predictive power.

References

[1] D. Batool, M. Shahbaz, H.S. Asif, K. Shaukat, T.M. Alam, I.A. Hameed, S. Luo, A hybrid approach to tea crop yield prediction using simulation models and machine learning. Plants, 11(15), (2022) 1925. https://doi.org/10.3390/plants11151925

[2] D. Dahiphale, P. Shinde, K. Patil, V. Dahiphale, (2023) Smart farming: Crop recommendation using machine learning with challenges and future ideas. Authorea Preprints. https://doi.org/10.36227/techrxiv.23504496.v1

[3] K.A. Reddy, R.K. Kumar, Recommendation System: A Collaborative Model for Agriculture. International Journal of Computer Sciences and Engineering, 6(1), (2018) 120-123. https://doi.org/10.26438/ijcse/v6i1.120123

[4] D. Femi and A. M. Mukunthan, Plant leaf infected spot segmentation using robust encoder-decoder cascaded deep learning model. Network: Computation in Neural Systems, (2023) 1–19. https://doi.org/10.1080/0954898X.2023.2286002

[5] K.O. McGraw, S.P. Wong, Forming inferences about some intraclass correlation coefficients. Psychological methods, 1(1), (1996) 30-46. https://psycnet.apa.org/doi/10.1037/1082-989X.1.1.30

[6] G. Adomavicius, A. Tuzhilin, Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge & Data Engineering, 17(6), (2005) 734-749. https://doi.org/10.1109/TKDE.2005.99

[7] Atharva Ingle, (2020) *Crop Recommendation Dataset* (Version V1). [Dataset]. Kaggle. https://www.kaggle.com/datasets/atharvaingle/crop-recommendation-dataset

[8] S. Gupta, A. Geetha, K. S. Sankaran, A.S. Zamani, M. Ritonga, R. Raj, H. S. Mohammed, Machine learning- and feature selection-enabled framework for accurate crop yield prediction. Journal of Food Quality, 2022(1), (2022) 6293985. https://doi.org/10.1155/2022/6293985

[9] H. Liu, H. Motoda, (2007) Computation method of feature selction, CRC Press, 440.

[10] M.L. Zhang, Z.H. Zhou, A review on multi-label learning algorithms. IEEE transactions on knowledge and data engineering, 26(8), (2013) 1819-1837. https://doi.org/10.1109/TKDE.2013.39

[11] S.N. Khan, D. Li, M. Maimaitijiang, Geographically weighted random forest approach to predict corn yield in the US Corn Belt. Remote Sensing, 14(12), (2022) 2843. https://doi.org/10.3390/rs14122843

[12] L. Zhang, L. Xie, Z. Wang, C. Huang, Cascade parallel random forest algorithm for predicting rice diseases in big data analysis. Electronics, 11(7), (2022) 1079. https://doi.org/10.3390/electronics11071079

[13] A. Bhullar, K. Nadeem, R.A. Ali, Simultaneous multi-crop land suitability prediction from remote sensing data using semi-supervised learning. Scientific Reports, 13(1), (2023) 6823. https://doi.org/10.1038/s41598-023-33840-6

[14] H. Burdett, C. Wellen, Statistical and machine learning methods for crop yield prediction in the context of precision agriculture. Precision Agriculture, 23, (2022) 1553–1574. https://doi.org/10.1007/s11119-022-09897-0

[15] D. Tang, Y. Liu, and T.-K. Kim, Fast pedestrian detection by cascaded random forest with dominant orientation templates. British Machine Vision Conference (BMVC), 2016.

[16] S. Janrao, K. Shah, A. Pavate, R. Patil, S. Bankar, A. Vasoya, Conglomerate crop recommendation by using multi-label learning via ensemble supervised clustering techniques. International Research Journal of Multidisciplinary Technovation, 6(3), (2024) 90–100. https://doi.org/10.54392/irjmt2437

[17] P.S. Kiran, G. Abhinaya, S. Sruti, N. Padhy, A Machine Learning-Enabled System for Crop Recommendation. Engineering Proceedings, 67(1), (2024) 51. https://doi.org/10.3390/engproc2024067051

[18] A. Maheswary, S. Nagendram, K.U. Kiran, S.H. Ahammad, P.P. Priya, M.A. Hossain, A.N.Z. Rashed, Intelligent Crop Recommender System for Yield Prediction Using Machine Learning Strategy. Journal of the Institution of Engineers (India): Series B, 105(4), (2024) 979-987. https://doi.org/10.1007/s40031-024-01029-8

[19] B. Dey, J. Ferdous, R. Ahmed, Machine learning based recommendation of agricultural and horticultural crop farming in India under the regime of NPK, soil pH and three climatic variables. Heliyon, 10(3), (2024). https://doi.org/10.1016/j.heliyon.2024.e25112

[20] X. Yuan, S. Liu, W. Feng, G. Dauphin, Feature importance ranking of random forest-based end-to-end learning algorithm. Remote Sensing, 15(21), (2023) 5203. https://doi.org/10.3390/rs15215203

[21] M.K. Senapaty, A. Ray, N. Padhy, IoT-enabled soil nutrient analysis and crop recommendation model for precision agriculture. Computers, 12(13), (2023) 61. https://doi.org/10.3390/computers12030061

[22] R.K. Rajak, A. Pawar, M. Pendke, P. Shinde, S. Rathod, A. Devare, Crop recommendation system to maximize crop yield using machine learning technique. International Research Journal of Engineering and Technology, 4(12), (2017) 950-953.

[23] S.K. Apat, J. Mishra, K.S. Raju, N. Padhy, An artificial intelligence-based crop recommendation system using machine learning. Journal of Scientific & Industrial Research (JSIR), 82(5), (2023) 558–567. https://doi.org/10.56042/jsir.v82i05.1092

[24] Y.S. Mahmoud, S.A. Gamel, F.M. Talaat, Enhancing crop recommendation systems with

explainable artificial intelligence: A study on agricultural decision-making. Neural Computing and Applications, 36(1), (2024) 5695–5714. https://doi.org/10.1007/s00521-023-09391-2

[25] M.K. Senapaty, A. Ray, N. Padhy, A decision support system for crop recommendation using machine learning classification algorithms. Agriculture, 14(8), (2024) 1256. https://doi.org/10.3390/agriculture14081256

[26] S. Rani, A.K. Mishra, A. Kataria, S. Mallik, H. Qin, Machine learning-based optimal crop selection system in smart agriculture. Scientific Reports, 13, (2023) 15997. https://doi.org/10.1038/s41598-023-42356-y

[27] S.S. Saritha, An experimental analysis of machine learning techniques for crop recommendation. Nigerian Journal of Technology, 43(2), (2024). https://doi.org/10.4314/njt.v43i2.13

[28] G.B. Dela Cruz, B.D. Gerardo, B.T. Tanguilig III, Agricultural Crops Classification Models Based on PCA-GA Implementation in Data Mining. International Journal of Modeling and Optimization, 4(5), (2014) 375. https://doi.org/10.7763/IJMO.2014.V4.404

[29] S.K.S. Durai, M.D. Shamili, Smart farming using Machine Learning and Deep Learning techniques. Decision Analytics Journal, 3, (2022) 100041. https://doi.org/10.1016/j.dajour.2022.100041

[30] B.F. Darst, K.C. Malecki, C.D. Engelman, Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genomic Data, 19(S1), (2018) 65. https://doi.org/10.1186/s12863-018-0633-8

Authors Contribution Statement

Umamaheswari R: Conceptualization, Methodology, Software, Formal analysis, Writing – original draft, Review of Existing Literatures, Paper Formatting. E. Kannan: Formal analysis, Review of work, Methodology analysis. All authors have read and agreed to the published version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Has this article screened for similarity?

Yes

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

About the License