Indian Journal of Agricultural Sciences **95** (11): 1324–1331, November 2025/Article
https://doi.org/10.56093/ijas.v95i11.166636

Explainable XGBoost model for crop recommendation in Mizoram using hybrid random forest and particle swarm optimization

ZAITINKHUMA THIHLUM[1], V D AMBETH KUMAR[1]* and A K MOHANTY[2]

Mizoram University, Tanhril, Aizawl, Mizoram 796004, India

Received: 14 May 2025; Accepted: 14 November 2025

ABSTRACT

The study was carried out during 2023–2024 using multi-location data across the three districts of Lawngtlai, Serchhip, and Champhai to support sustainable agriculture in Mizoram, North-east India, with the help of XGBoost-based crop recommendation system. A hybrid feature selection approach combining Random Forest (RF) and Particle Swarm Optimization (PSO) was proposed to identify key agronomic features. Class imbalance was addressed using Synthetic Minority Oversampling Technique (SMOTE), and model performance was evaluated using standard metrics such as accuracy, precision, recall, and F1-score. GridSearchCV was employed for hyperparameter optimization, with a 5-fold cross-validation applied to validate model performance during training. The XGBoost classifier trained on the hybrid RF + PSO-optimized features outperformed those trained on the full feature set and the RF top-8 features, owing to the combined benefits of SMOTE-based class balancing and PSO-driven optimal feature selection. The SHAP analysis for the four major crops rice, maize, moong, and potato revealed that nitrogen (N) and potassium (K) were the most influential factors shaping crop prediction outcomes, followed by phosphorus (P) and soil *p*H, while rainfall had the least influence due to Mizoram's consistently high and evenly distributed precipitation across its cultivation zones. The proposed approach enhances both accuracy and interpretability, providing a reliable decision-support framework for crop selection tailored to Mizoram's diverse agro-climatic conditions.

Keywords: Crop recommendation, Feature selection, Precision agriculture, RF-PSO, SHAP, XGBoost

Agriculture is vital to global food security, producing 23.7 million tonnes daily and supporting 2.5 billion livelihoods. In developing countries, it contributes 29% to GDP and employs 65% of the population (Convention on Biodiversity 2018). With the global population expected to exceed 9 billion by 2050, food production must rise by 60–70%. As reported by WHO over 820 million people face food shortages, prompting UN efforts to promote sustainable agriculture and double smallholder incomes by 2030 (Lee *et al.* 2016).

In India, agriculture contributes 16.5% to GDP and employs 42.3% of the workforce (Gulati and Juneja 2022). In Mizoram, over half the population practices *jhum*, a traditional slash-and-burn method despite considered unsustains due to environmental degradation, biodiversity loss, shorter cycles, and declining yields (Sati 2019, Pachuau and Devi 2020, Rohlupuii *et al.* 2023). Though wet rice and plantation farming are encouraged, *jhum* persists in remote areas due to limited resources (Thanga 2020). Productivity is tied to monsoon rainfall, peaking in August and lowest in January (Pandey *et al.* 2024). Enhancing *jhum* with better soil, water, and fire management could improve sustainability (Tripathi *et al.* 2024). Integrated farming systems (IFS) require greater investment (Kumar *et al.* 2023). The Mizoram Organic Mission and regional initiatives support certified clusters and organic markets (Singh *et al.* 2021, Darjee 2023).

Machine learning (ML) and artificial intelligence (AI) are being incorporated into precision agriculture to improve predictions and maximize resource utilization (Mor *et al*. 2021, Kumar *et al*. 2022). The RBF-SMO and PSO-MDNN methods have overcome the conventional methods in terms of accuracy (Mythili and Rangaraj 2021). Furthermore, ensemble ML models have demonstrated great potential in improving crop yield estimation (Hasan *et al*. 2023). Furthermore, crop recommendations based on AI can help maximize resource usage and encourage sustainable farming methods (Saxena *et al.* 2020, Thihlum *et al.* 2025). The main objective of the study is to design a precision farming decision-making system that incorporates improved feature selection, class balancing, explainable AI, and the best ML strategies to improve crop choice and productivity in Mizoram.

MATERIALS AND METHODS

This section presents various materials, including the dataset and preliminary operations on the input data for the

[1]Mizoram University, Tanhril, Aizawl, Mizoram; [2]ICAR-Agricultural Technology Application Research Institute, Umiam, Meghalaya. *Corresponding author email: ambeth@mzu.edu.in

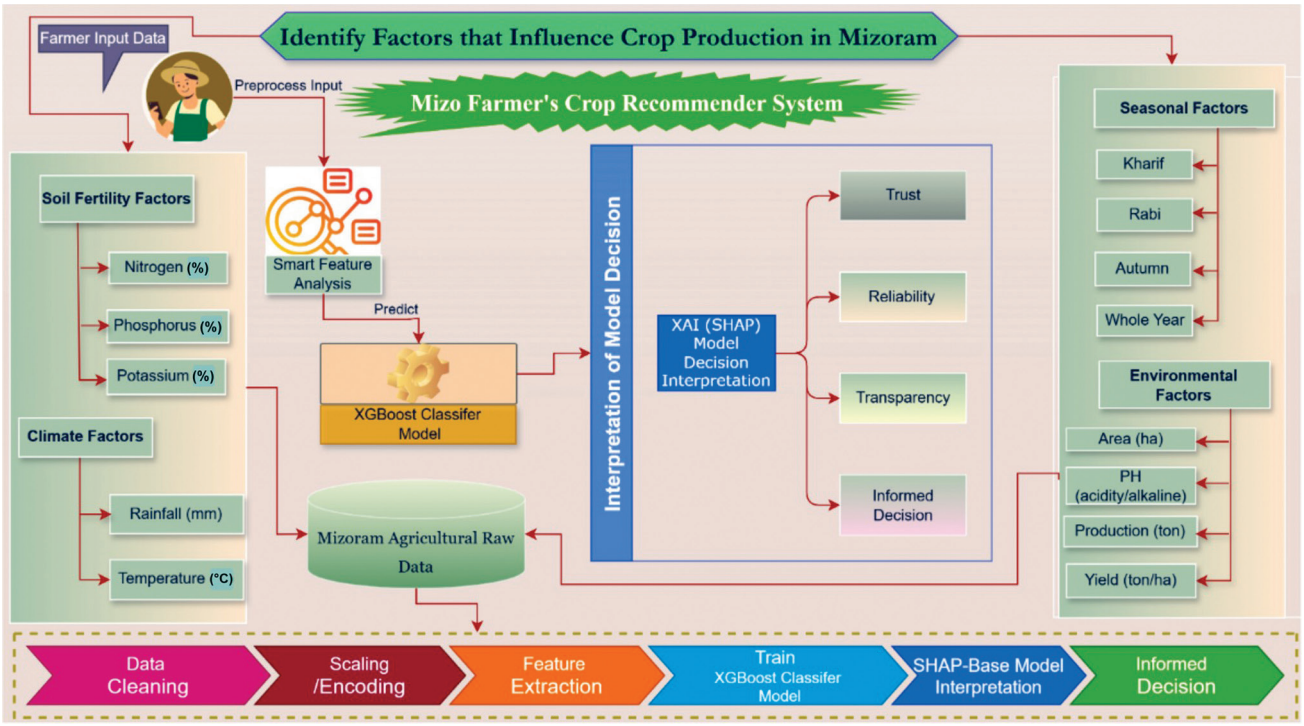

Fig. 1 Factors that influence crop production in Mizoram and the flow of the proposed crop recommender system to assist decision making for Mizoram farmers.

XGBoost model, feature extraction techniques, performance metrics, and SHAP analysis for model interpretation. The overall workflow of the study, illustrating factors influencing crop production in Mizoram and the flow of the proposed crop recommender system is shown in Fig. 1.

Data collection: The study was carried out during 2023–2024 in Mizoram, North-east India, covering three districts (Lawngtlai, Serchhip, and Champhai). These districts represent distinct agro-climatic zones and soil types, forming a multi-location trial framework. Data were obtained from the Mizoram Crop Recommendation Dataset (Pandey 2021) and seasonal rainfall data (mm) from the Meteorological Data of Mizoram (MDM) (https://des.mizoram.gov.in/), respectively. Crops with fewer than 10 valid records were excluded to ensure uniformity. Based on district-wise agricultural statistics for 2023–2024 (Department of Agriculture, Government of Mizoram 2024; https://agriculturemizoram.nic.in/), nine major crops shown in Fig. 2 were selected for analysis. The Lawngtlai district reported the following cultivation areas and production, rice 1,166 ha/3,491 Mt; maize 85 ha/340 Mt; potato 44 ha/183 Mt; cotton 7 ha/2 Mt; rapeseed 18 ha/4 Mt; sesamum 45 ha /13 Mt; soybean 68 ha/72 Mt; and pulses 68 ha/139 Mt. In Serchhip, the cultivated areas and production were, rice 705 ha/1,403 Mt; maize 396 ha/806 Mt; rapeseed 98 ha/59 Mt; sesamum 6 ha/5 Mt; soybean 11 ha/9 Mt, and pulses 39 ha /40 Mt. The Champhai district recorded, rice 2,495 ha/5,489 Mt; maize 1,238 ha/1,956 Mt, Potato 98 ha/829 Mt; rapeseed 70 ha/87 Mt, sesamum 365 ha/228 Mt; soybean 430 ha/539 Mt, and pulses 242 ha/348 Mt. These data reflect the cropping patterns and production diversity across Mizoram's hilly terrain, emphasizing the influence of local micro-climates and soil variability on crop productivity under a multi-location, single-year trial.

Data preprocessing: The selected crops were considered for the target classes, with numerical features including N (%), P (%), K (%), *p*H, Rainfall(mm), Temperature (°C), Area (ha), Production (tonnes), Yield (tonnes/ha) and one categorical feature season. All numerical features were scaled to a range of 0–1 using Min-Max scaling, ensuring uniform contribution of all features during model training. Categorical feature was encoded using one-hot encoding, where each category was represented as a binary vector, with 1 indicating the presence of a category and 0 otherwise. The pre-processed numerical and categorical features were

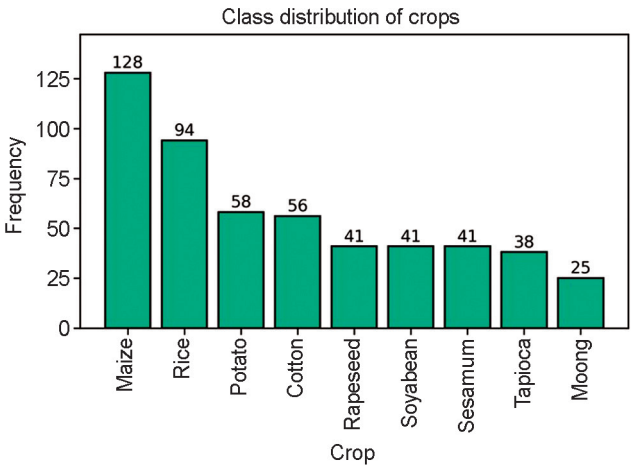

Fig. 2 Distribution of crop classes in the dataset.

combined into a single feature vector for each record. The dataset was subsequently split into training (80%) and testing (20%) sets. The preprocessed dataset was then used as input to an RF model for feature importance selection (De *et al*. 2023).

Impurity based feature selection using Random Forest (RF): RF is an ensemble learning method that builds multiple decision trees during training and aggregates their predictions for the final output (Elavarasan *et al.* 2020, Manokaran and Vairavel 2023). To determine feature importance, RF uses Gini impurity, which measures node purity. For a node j, the Gini impurity G_j is computed using Equation 1:

$$G_j = 1 - \sum_{i=1}^{c} p_{i,j}^2 \quad (1)$$

Where $P_{i,j}$ is the probability of class i at node j, and c is the number of classes (Nembrini *et al.* 2018). When a node is split, the reduction in impurity (nI_j) is calculated as the difference between the weighted impurity of the parent node and the sum of weighted impurities of the child nodes (Equation 2).

$$nI_j = w_j G_j - w_{jleft} G_{jleft} - w_{jright} G_{jright} \quad (2)$$

Where w_j is the proportion of samples at node j relative to the total dataset. The importance of a feature is computed by summing the impurity reductions for all splits involving that feature and then normalizing across all features. Finally, the RF feature importance is averaged across all trees in the forest. Features are ranked by their importance scores, and the top 8 features are selected for further analysis (Fig. 3A).

Particle Swarm Optimization (PSO) for feature subset optimization: The PSO introduced by Kennedy and Eberhart (1995), is a population-based global optimization algorithm inspired by bird flocking. Its effectiveness has been demonstrated in various domains (Elsheikh and Elaziz 2019, Mansouri *et al.* 2022). PSO is well-suited for feature selection due to its binary encoding simplicity, global search capability, computational efficiency, few parameters, and ease of implementation (Jain *et al.* 2022).

Let d be the total number of features in the dataset and $\subseteq d$ be the top 8 features selected via RF. PSO is applied to d' to identify the optimal subset, reducing the search space. This RF+PSO hybrid approach improves model accuracy by selecting the most impactful features (Fig. 3B). Logistic Regression (LR) is used to evaluate subsets during selection, with 5-fold cross-validation guiding the threshold choice (Qasim and Algamal 2018). The ideal threshold (i.e. θ = 0.14) value obtained using mean cross-validation accuracy with a 5-folds using the given range of 0.01–0.3. Equation 6 is used to compute the velocity and Equation 7 for position of the particle i. The Pseudo code for the PSO algorithm is described in the *Algorithm* 1.

$$\tilde{x}_i = [\tilde{x}_1, \tilde{x}_2, \ldots, \tilde{x}_{d'}]; \tilde{x}_i \in [0,1]^{d'} \quad (3)$$

$$\tilde{u}_i = [\tilde{u}_1, \tilde{u}_2, \ldots, \tilde{u}_{d'}] \quad (4)$$

$$\tilde{x}_i = \begin{cases} 1, \text{ if } \tilde{x}_i \theta, \\ 0, \text{ otherwise} \end{cases} \quad (5)$$

Data Balancing Using SMOTE: Data imbalance occurs when two or more class distributions are not equal. In classification problems, when classes are imbalanced, resampling becomes necessary since most ML algorithms' performance poorly with considerable class distribution differences and become biased toward the majority class (Chabalala *et al.* 2023). Other consequence can be poor generalization and sub-optimal performance related to the minority class, which in many cases can be much more important in real-world applications (Wang *et al.* 2021). SMOTE is utilized to address the aforementioned issues caused by the class imbalance. Given a sample $\tilde{x}_i$, an artificial sample (new data point) will be generated by randomly selecting its K-nearest neighbours. The K value is set to 5 neighbours (Li *et al.* 2021). Among these five neighbours, one neighbour $\tilde{x}_{zi}$ is selected and the new sample $\tilde{x}_{new}$ will be generated by computing the interpolation of $\tilde{x}_i$ and $\tilde{x}_{zi}$ using Equation 8.

$$\tilde{x}_{new} = \tilde{x}_i + \delta(\tilde{x}_{zi} - \tilde{x}_i) \quad (8)$$

Where δ is the random number in the range [0,1].

Model training: XGBoost is a scalable gradient boosting system widely used for classification and regression tasks. As an ensemble method, it combines multiple weak learners to generate precise and generalized predictions. Its efficiency and performance on large datasets have made it a popular

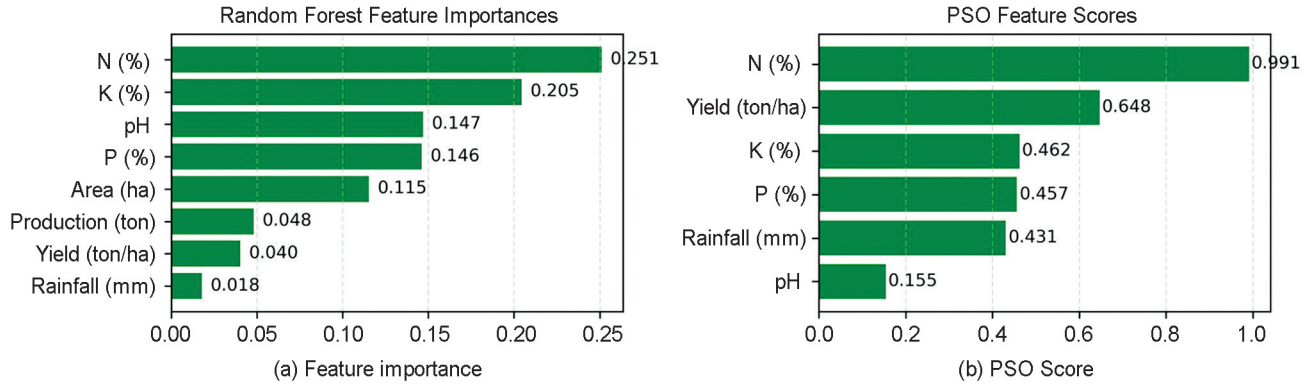

Fig. 3 (a) Illustrates the top 8 important features selected by Rf and, (b) Represents the PSO applied to RF sub-features for optimizing the feature selection process.

choice in machine learning competitions: in 2015, 17 out of 29 winning Kaggle solutions used XGBoost, with 8 exclusively relying on it, and all top 10 teams in the KDD Cup 2015 employed it (Vijay *et al*. 2021, Asselman *et al.* 2023). Conventional gradient boosting decision trees rely only on first-order derivatives and face challenges in parallel training due to dependencies among weak learners (Amjad *et al.* 2022). XGBoost overcomes this limitation by using a second-order Taylor expansion along with a regularization term, balancing loss reduction and model complexity to prevent overfitting (Li *et al.* 2023). Its tree structures are optimized using gradient and Hessian information, allowing efficient recursive partitioning. In this study, XGBoost was trained on the top eight features selected by RF (d') and the optimal subset chosen via PSO (d''). Seven key hyperparameters were tuned colsample_bytree, gamma, learning_rate, max_depth, min_child_weight, n_estimators, and subsample to control model complexity and improve performance (Garg and Alam 2023, Geng *et al.* 2023, Clarke *et al.* 2024). For a given input feature vector $\tilde{x}_i$, the predicted value for the i^{th} instance is obtained by aggregating the predictions of all trees in the ensemble (Equation 9).

$$\hat{y}_i = \sum_{t=1}^{\tau} f_\tau(\tilde{x}_i), f_t \in \Psi \quad (9)$$

Where f_τ is the function of the τ^{-th} decision tree, T is the total number of trees, and ψ represents the set of all possible trees (CART). XGBoost's objective function combines the loss function with a regularization term to control model complexity and prevent overfitting (Lv *et al.* 2021). Node splitting and leaf weights are determined to maximize gain while minimizing the objective function (Coffie and Cudjoe 2024).

K-fold cross-validation: During training, 5-fold cross-validation was employed to evaluate the model's generalization ability and to detect potential overfitting (Zhang and Liu 2023). The dataset was divided into five equal folds, with four folds used for training and the remaining fold for testing. This process was repeated for

Algorithm 1: Hybrid RF + PSO Algorithm for Feature Selection

Input: *Top 8 features selected by RF* ($\tilde{x}_i$), n_p =10, t = 5. // *Where* n_p *is swarm size, t is number of iterations*

Output: *Optimal subset of features*

Initialize Swarm:

Create a swarm with n_p particles

for *each particle i* ***do***

Set random initial position $\tilde{x}_i$ *and velocity* u_i *within bounds.*

End for

Evaluate Fitness:

for *each particle i* ***do***

Compute fitness $F(\tilde{x}_i)$ *using feature* $\tilde{x}_i$'

If $F(\tilde{x}_i) > (pbest_i)$ ***then*** // *Where* $pbest_i$ *is the best solution found by an individual particle*

Update personal best: $pbest_i \leftarrow \tilde{x}_i$

End if

if $F(pbest_i) > \underline{F}(gbest_i)$ ***then*** // *Where* $gbest_i$ *is the best solution found by the entire swarm*

Update global best: $(gbest_i) \leftarrow pbest_i$

End if

End for

Update Velocities and Positions:

for *each particle i* ***do***

Update velocity: $u_i^{t+1} = \gamma.u_i^t + \alpha_1.\rho_1(pbest_i) - x_i^t) + \alpha_2.\rho_2.(gbest_i - x_i^t)$ (6)

// *Where* γ *is inertia weight,* α_1 *and* α_2 *are acceleration coefficients, while* ρ_1, *and* ρ_2 *are random values in [0, 1].*

Update position: $x^{t+1} = x_i^t + u_i^{t+1}$ (7)

End for

Check Termination:

if *stopping criteria are met (e.g., maximum iterations or convergence)* ***then,***

End the algorithm

else

Repeat from Steps 2 and 4

End if

Return the global best position gbest and its fitness value F(gbest\ ) // the optimal subset of features.

each fold, and the results were averaged to obtain a robust estimate of model performance.

Performance evaluation metrics: The study used various performance metrics such as accuracy, precision, recall, and F1-score to evaluate the quality of the prediction made by the model (Ajayi *et al.* 2023). Accuracy (Equation 10) measures the overall correctness of the model prediction by calculating the ratio of correct predictions (i.e. TruePost and TrueNeg) to the total number of predictions. Precision (Equation 11) measures the ratio of TruePost to the sum of TruePost and FalsePost. Recall (Equation 12) measures the ability of the model to identify the actual positive by computing the ratio of TruePost to the sum of both TruePost and FalseNeg. F1-score (Equation 13) combined the harmonic mean of precision and recall into a single metric, thereby providing a balanced view of the model performance.

$$\text{Accuracy} = \frac{\text{TruePost+TrueNeg}}{\text{TruePost+TrueNeg+FalsePost+FalseNeg}} \quad (10)$$

Where, TruePost represents the number of examples correctly classified as positive, TrueNeg indicates the number of examples correctly classified as negative, FalsePost determines the number of incorrectly identified positive examples, and FalseNeg corresponds to the number of examples incorrectly predicted as negative by the model.

$$\text{Precision} = \frac{\text{TruePost}}{\text{TruePost + FalsePost}} \quad (11)$$

$$\text{Recall} = \frac{\text{TruePost}}{\text{TruePost + FalseNeg}} \quad (12)$$

$$\text{F1 – Score} = \frac{2 \times (\text{Precision} \times \text{Recall})}{\text{Precision + Recall}} \quad (13)$$

SHAP based analysis of the model decision: SHAP (SHapley Additive exPlanations) is a post-hoc, model-agnostic technique used to quantify the contribution of individual features to model predictions (Naga *et al.* 2024). For tree-based models, SHAP values are efficiently computed using Tree SHAP (Wang *et al.* 2022) as defined in Equation 1.

$$\phi_j(f, \tilde{x}) = \frac{1}{T}\sum_{\hat{o}=1}^{T}\left[f_i(\tilde{x}) - \mathbb{E}\left[f_i(\tilde{x}) \mid x \right] \right] \quad (14)$$

Where T is the number of trees, $f_\tau(\tilde{x}_i)$ is the prediction from the τ^{th} tree for instance $\tilde{x}_i$, and $\mathbb{E}[f_{\hat{o}}(\tilde{x}_i) \mid x_j]$ is the expected prediction considering feature *j*. Feature contributions are visualized using beeswarm plots, which display the SHAP values for all features across instances. Fig. 4 shows the SHAP beeswarm plots for the RF–PSO–selected features, obtained from the XGBoost model optimized using GridSearchCV.

RESULTS AND DISCUSSION

The performance of the proposed hybrid RF + PSO feature selection approach was evaluated using standard metrics, including accuracy, precision, recall, and F1-score.

Table 1 Comparison of Full features, RF Top 8 features Selected and Hybrid RF+PSO

Metrics	Full features				RF top 8				Hybrid RF+PSO			
	Train	Test	CV	CV (Std)	Train	Test	CV	CV (Std)	Train	Test	CV	CV (Std)
Accuracy	0.95	0.95	0.96	0.01	0.99	0.94	0.99	0	1	0.98	0.99	0.02
Recall	0.95	0.94	0.96	0.01	0.99	0.94	0.99	0	1	0.98	0.99	0.02
Precision	0.95	0.95	0.96	0.02	0.99	0.95	0.99	0	1	0.98	0.99	0.01
F1 Score	0.95	0.95	0.96	0.02	0.99	0.94	0.99	0	1	0.98	0.99	0.02

Table 2 Class-wise test performance for full features, top 8 features selected by RF and Hybrid RF+PSO

Class	Precision			Recall			F1-Score		
	Full	RF	Hybrid RF+PSO	Full	RF	Hybrid RF+PSO	Full	RF	Hybrid RF+PSO
Cotton	0.86	1	0.91	0.93	1	1	0.89	1	0.95
Maize	0.93	0.92	0.96	0.95	0.89	0.96	0.92	0.91	0.96
Moong	0.95	0.67	1	0.93	0.8	1	0.94	0.8	1
Potato	0.93	1	1	0.95	1	1	0.96	1	1
Rapeseed	0.93	1	1	0.81	1	0.89	0.87	1	0.94
Rice	0.93	0.95	1	0.95	0.91	1	0.95	0.93	1
Sesamum	0.93	1	1	0.92	1	1	0.95	1	1
Soyabean	0.94	1	1	0.96	0.67	1	0.96	0.8	1
Tapioca	0.95	1	1	0.93	1	1	0.94	1	1

Full = Full features, RF = RF-selected top 8 features, RF+PSO = Model trained on RF top 8 features + PSO-selected features

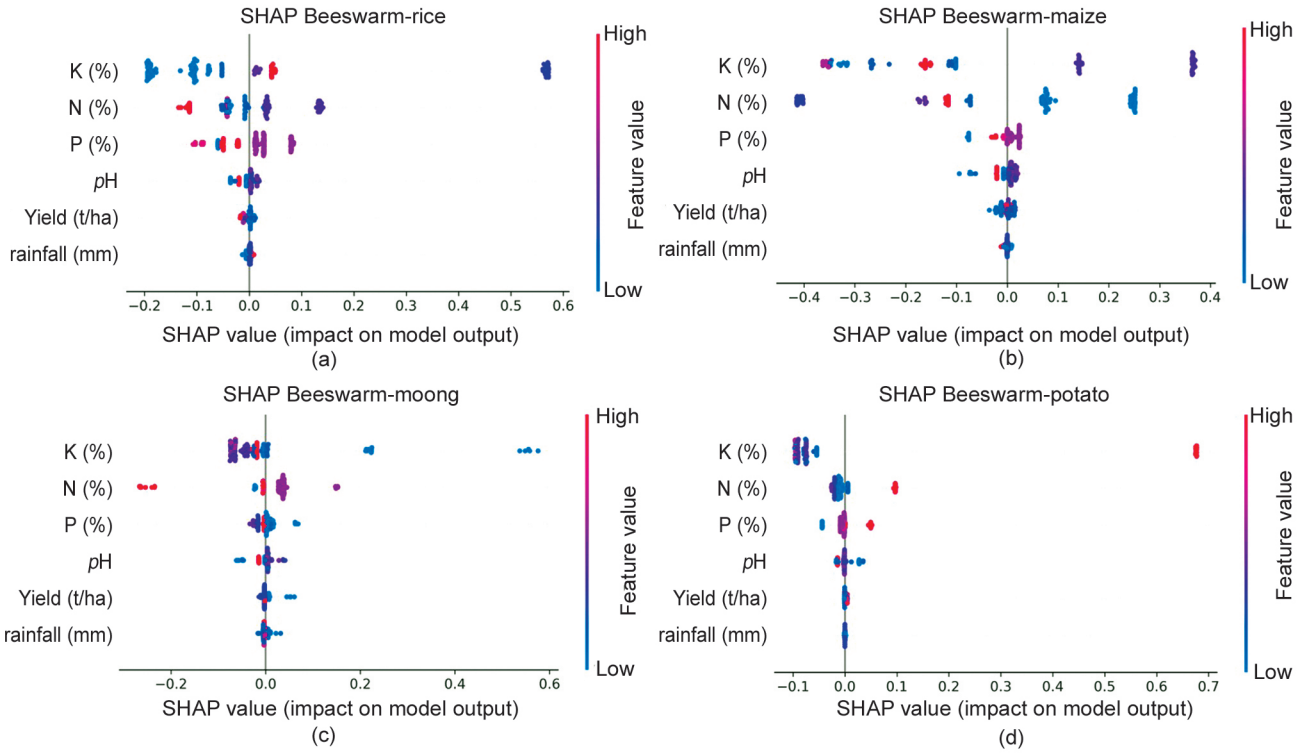

Fig. 4 SHAP beeswarm plots illustrating the impact of RF + PSO–selected features for (a) rice, (b) maize, (c) moong, and (d) potato classes.

Experiments were conducted on three feature sets, full features, the top 8 features selected by RF, and the optimized hybrid RF + PSO subset. The XGBoost model was fine-tuned using GridSearchCV with a consistent hyperparameter search space (Supplementary Table 1), and the optimal parameters were employed for final model training. Table 1 presents a comprehensive comparison of model performance across the three feature sets, while the corresponding baseline results are detailed in Supplementary Table 2, where the hybrid RF + PSO feature subset outperformed all other configurations including the full feature set, Decision Tree (DT) Top-8, and RF Top-8 across all evaluated metrics. The cross-validation mean and standard deviation values further ensure model stability and reduced overfitting achieved through hybrid optimization. The learning convergence of the tuned hybrid model is illustrated in Supplementary Fig. 1, which demonstrates consistent alignment between training and validation curves across the metrics, indicating stable model learning and reliable generalization behavior. Comparable findings have been reported in recent studies (Supplementary Table 3), where PSO-based hybrid models enhanced crop yield prediction and classification accuracy through global search optimization (Akbari *et al.* 2020, Mythili and Rangaraj 2021, Naga *et al.* 2024, Pan and Chen 2024, Xu and Sun 2025). Subsequently, SHAP analysis was performed for the rice, maize, moong and potato classes to interpret feature contributions and provide transparency to model predictions outcome. This interpretability analysis complements the quantitative evaluation in Table 1, offering deeper insights into how feature optimization impacts prediction outcome. Moreover, the precision-recall (PR) curve (Supplementary Fig. 2) further validates the robustness and discriminative capability of the proposed hybrid RF + PSO XGBoost model across multiple crop classes.

SHAP analysis of hybrid RF + PSO selected features for major crops (rice, maize, moong and potato): The SHAP analysis was employed to interpret the prediction behavior of the hybrid RF + PSO + XGBoost model across four major crops such as rice, maize, moong, and potato (Fig. 4). The results indicated that nitrogen (N) and potassium (K) were the most influential features driving yield predictions across all crop types, reflecting their critical roles in nutrient uptake and crop growth. For rice and maize, higher N and K levels contributed positively to model output, while rainfall and *p*H showed moderate influence. For moong and potato, temperature-linked soil fertility attributes, particularly N and P, exerted a stronger effect on prediction accuracy. Overall, the SHAP interpretation underscores the framework's ability to capture biologically meaningful relationships between nutrient composition, climatic factors, and crop suitability across Mizoram's diverse agro-ecological zones.

The proposed crop recommender system facilitates the farmer in selecting the crops best suited for the specific soil properties and local climatic conditions, which could potentially increase the agricultural output. Furthermore, providing alternatives to *Jhum* production encourages settled, stable farming systems while lowering soil degradation and increasing yield stability. The identification of critical factors such as nutrient content (nitrogen and potassium), *p*H, and land productivity helps policymakers prioritize interventions for soil health, input optimization, and regional crop choices. Furthermore, effective model categorization, particularly

for minority classes using SMOTE, enables reliable crop output estimation, enhancing supply chain planning and minimizing post-harvest losses, resulting in a more robust and responsive food system. However, as this study uses a single-year dataset from three districts, its generalizability may be limited; future work will include multi-year, multi-location data to enhance model robustness and applicability.

FUNDING SOURCE OF RESEARCH

This research was supported under the Tribal Area Sub Plan (TSP) scheme of IBITF-DST (Ref. No.: IBITF/Note/TSP/SanctionLetter/2023-24/0252).

REFERENCES

Ajayi O G, Ashi J and Guda B. 2023. Performance evaluation of YOLO v5 model for automatic crop and weed classification on UAV images. *Smart Agricultural Technology* **5**: 100231.

Akbari E, Darvishi Boloorani A, Neysani Samany N, Hamzeh S, Soufizadeh S and Pignatti S. 2020. Crop mapping using Random Forest and Particle Swarm Optimization based on multi-temporal Sentinel-2. *Remote Sensing* **12**(9): 1449.

Amjad M, Ahmad I, Ahmad M, Wróblewski P, Kaminski P and Amjad U. 2022. Prediction of pile bearing capacity using XGBoost algorithm: Modeling and performance evaluation. *Applied Sciences* **12**(4): 2126.

Asselman A, Khaldi M and Aammou S. 2023. Enhancing the prediction of student performance based on the machine learning XGBoost algorithm. *Interactive Learning Environments* **31**(6): 3360–79.

Chabalala Y, Adam E and Ali K A. 2023. Exploring the effect of balanced and imbalanced multi-class distribution data and sampling techniques on fruit-tree crop classification using different machine learning classifiers. *Geomatics* **3**(1): 70–92.

Clarke A, Yates D, Blanchard C, Islam M Z, Ford R, Rehman S and Walsh R. 2024. The effect of dataset construction and data pre-processing on the eXtreme Gradient Boosting algorithm applied to head rice yield prediction in Australia. *Computers and Electronics in Agriculture* **219**: 108716.

Coffie G H and Cudjoe S K. 2024. Using extreme gradient boosting (XGBoost) machine learning to predict construction cost overruns. *International Journal of Construction Management* **24**(16): 1742–50.

Convention on Biological Diversity. 2018. 2.6 billion people draw their livelihoods mostly from agriculture. Convention on Biological Diversity, Montreal.

Darjee D K. 2023. A comparative review and analysis of organic farming policies adopted by the north-east states of India: An exploratory study. *Journal of Emerging Technologies and Innovative Research* **10**(12): h555–h570.

De Amorim L B, Cavalcanti G D and Cruz R M. 2023. The choice of scaling technique matters for classification performance. *Applied Soft Computing* **133**: 109924.

Elavarasan D, Vincent P M D R, Srinivasan K and Chang C Y. 2020. A hybrid CFS filter and RF-RFE wrapper-based feature extraction for enhanced agricultural crop yield prediction modeling. *Agriculture* **10**(9): 400.

Elsheikh A H and Abd Elaziz M. 2019. Review on applications of particle swarm optimization in solar energy systems. *International Journal of Environmental Science and Technology* **16**: 1159–70.

Garg D and Alam M. 2023. An effective crop recommendation method using machine learning techniques. *International Journal of Advanced Technology and Engineering Exploration* **10**(102): 498.

Geng X, Wu S, Zhang Y, Sun J, Cheng H, Zhang Z and Pu S. 2023. Developing hybrid XGBoost model integrated with entropy weight and Bayesian optimization for predicting tunnel squeezing intensity. *Natural Hazards* **119**(1): 751–71.

Gulati A and Juneja R. 2022. Transforming Indian Agriculture. *Indian Agriculture Towards 2030*, pp. 9–37. Chand R, Joshi P and Khadka S (Eds). Springer, Singapore.

Hasan M, Marjan M A, Uddin M P, Afjal M I, Kardy S, Ma S and Nam Y. 2023. Ensemble machine learning-based recommendation system for effective prediction of suitable agricultural crop cultivation. *Frontiers in Plant Science* **14**: 1234555.

Jain M, Saihjpal V, Singh N and Singh S B. 2022. An overview of variants and advancements of PSO algorithm. *Applied Sciences* **12**(17): 8392.

Rohlupuii, Kaur A, Kataria P and Laishram P. 2023. Assessment of crop production dynamics in Mizoram. *Agricultural Reviews* **44**(4): 573–76.

Kennedy J and Eberhart R. 1995. Particle swarm optimization. *(In) Proceedings of ICNN'95-International Conference on Neural Networks*, Perth, Western Australia, Australia, 27 November–1 December, pp. 1942–48.

Kumar M, Maurya P and Verma R. 2022. Future of Indian Agriculture Using AI and Machine Learning Tools and Techniques. *The New Advanced Society: Artificial Intelligence and Industrial Internet of Things Paradigm*, pp. 447–72. Panda S K, Mohapatra R K, Panda S and Balamurugan S (Eds). Scrivener Publishing, Beverly.

Kumar Y B, Lalramhlimi B, Lalrinsanga P L, Soni J K and Doley S. 2023. Success of integrated farming system for enhancing farmer's income in Mizoram. *Indian Farming* **73**(8): 39–43.

Lee B X, Kjaerulf F, Turner S, Cohen L, Donnelly P D, Muggah R, Davis R, Realini A, Kieselbach B, MacGregor L S and Waller I. 2016. Transforming our world: Implementing the 2030 agenda through sustainable development goal indicators. *Journal of Public Health Policy* **37**(Suppl 1): 13–31.

Li J, Zhu Q, Wu Q and Fan Z. 2021. A novel oversampling technique for class-imbalanced learning based on SMOTE and natural neighbors. *Information Sciences* **565**: 438–55.

Li Y, Zeng H, Zhang M, Wu B, Zhao Y, Yao X, Cheng T, Qin X and Wu F. 2023. A county-level soybean yield prediction framework coupled with XGBoost and multidimensional feature engineering. *International Journal of Applied Earth Observation and Geoinformation* **118**:103269.

Lv C X, An S Y, Qiao B J and Wu W. 2021. Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model. *BMC Infectious Diseases* **21**: 1–3.

Manokaran J and Vairavel G. 2023. GIWRF-SMOTE: Gini impurity-based weighted random forest with SMOTE for effective malware attack and anomaly detection in IoT-Edge. *Smart Science* **11**(2): 276–92.

Mansouri M, Safavi H R and Rezaei F. 2022. An improved MOPSO algorithm for multi-objective optimization of reservoir operation under climate change. *Environmental Monitoring and Assessment* **194**(4): 261.

Mor S, Madan S and Prasad K D. 2021. Artificial intelligence and carbon footprints: Roadmap for Indian agriculture. *Strategic Change* **30**(3): 269–80.

Mythili K and Rangaraj R. 2021. Deep learning with particle swarm based hyper parameter tuning based crop recommendation for better crop yield for precision agriculture. *Indian Journal of Science and Technology* **14**(17): 1325–37.

Naga S P, Ijaz M F and Woźniak M. 2024. XAI-driven model for crop recommender system for use in precision agriculture. *Computational Intelligence* **40**(1): e12629.

Nembrini S, Konig Ira and Wright M N. 2018. The revival of the Gini importance? *Bioinformatics* **34**(21): 3711–18.

Pachuau A L and Devi O H. 2020. Shifting cultivation and environment. A study of subsistence to profit in Mizoram. *Mizoram University Journal of Humanities and Social Science* **6**(1): 47–61.

Pandey A. 2021. *Crop Production in India*. Kaggle Dataset. Available at: https://www.kaggle.com/datasets/asishpandey/crop-production-in-india. Accessed on August 27, 2024.

Pandey V, Pandey P K, Chakma B and Ranjan P. 2024. Influence of short-and long-term persistence on identification of rainfall temporal trends using different versions of the Mann-Kendall test in Mizoram, north-east India. *Environmental Science and Pollution Research* **31**(7): 10359–378.

Pan X and Chen J. 2024. The optimization path of agricultural industry structure and intelligent transformation by deep learning. *Scientific Reports* **14**: 29548.

Qasim O S and Algamal Z Y. 2018. Feature selection using particle swarm optimization-based logistic regression model. *Chemometrics and Intelligent Laboratory Systems* **182**: 41–46.

Sati V P. 2019. Shifting cultivation in Mizoram, India. An empirical 2020 study of its economic implications. *Journal of Mountain Science* **16**: 2136–49.

Saxena A, Suna T and Saha D. 2020. Application of artificial intelligence in Indian agriculture. *(In) Souvenir: 19 National Convention-Artificial Intelligence in Agriculture: Indian perspective*. RCA Alumni Association, May. Udaipur, pp. 14-22.

Singh R, Babu S, Avasthe R, Das A, Praharaj C, Layek J, Kumar A, Rathore S, Mrunalini K, Kumar S, Yadav S and Pashte V. 2021. Organic farming in North-East India: Status and strategies. *Indian Journal of Agronomy* **66**(5): 163–79.

Thanga J L. 2020. Land use policies in the state of Mizoram. *Journal of Economic and Social Development* **16(**1–2**)***:* 4483.

Thihlum Z, Ambeth Kumar V D and Chawngsangpuii. 2025. Impact of SMOGN on regression models for crop yield prediction in Mizoram agriculture. *(In) Proceedings of International Conference on Soft Computing and its Engineering Applications*, pp. 170–182. Patel K K, Santosh K C, Oliveira G G de, Patel A and Ghosh A (Eds). Springer, Cham.

Tripathi S K, Hauchhum R, Ovung E Y, Singh N S, Vanlalfakawma D C, Upadhyay K K, Brearley F Q and Lalraminghlova H. 2024. Innovative shifting cultivation and other agricultural practices conducted by the indigenous population of Mizoram, north-east India. *Shifting Cultivation Systems*, pp. 29–48. Tripathi S K and Brearley F Q (Eds). Springer, Cham.

Vijay R, Manoj S, Ravikanth V, Vikas Y and Priyadarshini P I. 2021. Augmenting network intrusion detection system using extreme gradient boosting (XGBoost). *International Journal of Creative Research Thoughts* **9**(6): b550–b556.

Wang S, Dai Y, Shen J and Xuan J. 2021. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports* **11**(1): 24039.

Wang D, Thunéll S, Lindberg U, Jiang L, Trygg J and Tysklind M. 2022. Towards better process management in wastewater treatment plants: Process analytics based on SHAP values for tree-based machine learning methods. *Journal of Environmental Management* **301**: 113941.

Xu Z and Sun Y. 2025. Particle swarm optimization for agricultural problems. *Highlights in Business, Economics and Management* **51**: 245–50.

Zhang X and Liu C A. 2023. Model averaging prediction by K-fold cross-validation. *Journal of Econometrics* **235**(1): 280–301.