IJFANS INTERNATIONAL JOURNAL OF FOOD AND NUTRITIONAL SCIENCES
ISSN PRINT 2319 1775 Online 2320 7876
 Journal UGC CARE Listed (Group-I) Volume 12 , Issu 01 2023

INTEGRATING STACKING ENSEMBLE LEARNING FOR SOIL ANALYSIS AND CROP PREDICTION USING OPTIMIZED FEATURE SELECTION APPROACH

G. Uma[1], Dr. V. Joseph Peter[2]

[1]Research Scholar, [2] Associate Professor, [1]Reg No: 22122102302005

guma1728@gmail.com

[1,2] Research Department of Computer Science, Kamaraj College, Thoothukudi, Affiliated to Manonmaniam Sundaranar University, Abishekapatti,

Tirunelveli - 627012, Tamil Nadu, India.

peterchristy_in@yahoo.com

Received: 02.01.2023, Reviewed: 07.01.2023, Accepted: 19.01.2023, Published: 28.01.2023

Abstract:

Accurate and efficient crop prediction is crucial for optimizing agricultural productivity and supporting sustainable farming practices. In precision agriculture, accurate crop recommendations are vital for improving agricultural productivity and sustainability. This research proposes an advanced crop recommendation system that integrates Stacking Ensemble Learning with an Optimized Feature Selection approach for enhanced soil analysis and crop prediction. Additionally, an optimized feature selection technique is applied to identify the most relevant features from large agricultural datasets, enhancing model performance by eliminating irrelevant or redundant data. This work utilizes soil properties, climate conditions, and historical crop data to recommend the most suitable crops for specific environments. In this paper, the feature selection techniques namely Recursive Feature Elimination (RFE) and Boruta are used to select the features. Integrating Stacking Ensemble Learning (ISEL) technique consists of k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Tree (DT) algorithms to classify the classes. The Meta classifier SVM is used to predict and recommend the suitable crops to the farmers.

Keywords: Recursive Feature Elimination (RFE), Boruta, Integrating Stacking Ensemble Learning (ISEL), k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Tree (DT).

1. INTRODUCTION

Predicting suitable crops for cultivation is an important part of agriculture, and machine learning algorithms have played a significant role in this process in recent years. In this age of technology and data science, the agricultural sector can benefit greatly from properly implemented techniques. Feature selection and classification are important machine learning techniques. Feature selection involves selecting the most important attributes from a dataset. It involves selecting a subset of appropriate attributes from a larger set of original attributes based

on a predefined benchmark, such as classification performance or class separability, which is important in machine learning applications. The attribute selection uses three feature selection methods: filter, wrapper, and embedded. Filter methods execute quickly, whereas wrapper methods have a higher recognition rate. This work predicts and recommends suitable crops to farmers based on the real time Manur block soil dataset and the early cultivation [1]. In this study, wrapper feature selection techniques are used to select the best attributes from the dataset, followed by classification to predict the best crop for a specific piece of land based on the selected attributes. There are three types of machine learning techniques: supervised, unsupervised, and reinforcement learning. This work employs supervised learning classification techniques for prediction. The primary contribution of this work is to identify the best feature selection technique, combined with a classification method, to predict the most suitable crop for cultivation based on factors such as soil and environment.

The proposed system consists of three different processes namely preprocessing, feature selection and prediction and recommendation of crop yielding to farmers from actual dataset. The crop wise dataset is taken for predicting the recommended crop using soil data. The system employs multiple base models, including decision trees, support vector machines, and neural networks, which are combined to form a robust ensemble model that improves prediction accuracy and reduces overfitting. The feature selection techniques namely Recursive Feature Elimination (RFE) and Boruta [2] are used to select the features. Integrating Stacking Ensemble Learning (ISEL) technique [3] consists of k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Tree (DT) algorithms to classify the classes. The Meta classifier SVM is used to predict and recommend the suitable crops to the farmers. Fig. 1 shows the block diagram of the proposed work.

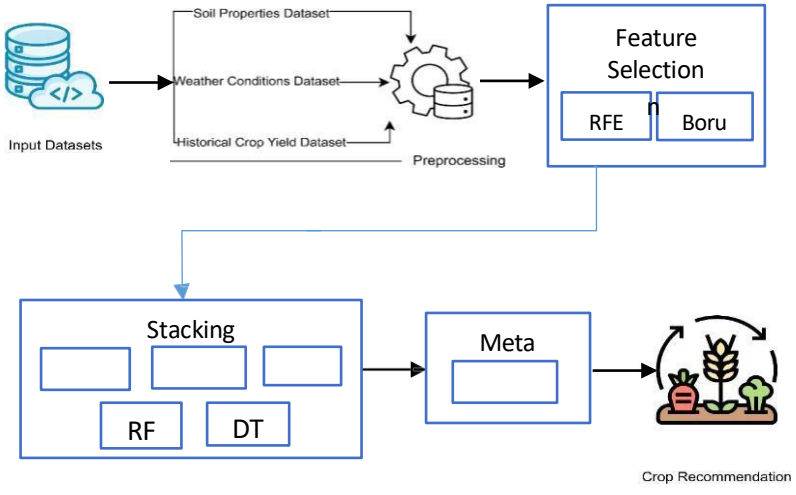

Fig. 1 Proposed Framework of ISEL Model

2. LITERATURE REVIEW

Recent advancements in machine learning (ML) and artificial intelligence (AI) have significantly transformed agricultural practices, offering data-driven insights for precision farming, soil analysis, and yield prediction. Researchers have explored various ML algorithms

to analyze soil properties, optimize resource utilization, and enhance crop productivity. Traditional statistical methods have been widely used in agricultural modeling, but ML-based approaches offer superior predictive accuracy and adaptability to complex, nonlinear relationships in agricultural data.

Machine learning (ML) is an important decision support tool for crop yield prediction, including supporting decisions on what crops to grow and what to do during the growing season of the crops. [4] Machine learning can make use of agricultural data related to crop yield under varying soil nutrient levels, and climatic fluctuations to suggest appropriate crops or supplementary nutrients to achieve the highest possible production [5].

Machine learning (ML) techniques improve soil fertility prediction by analyzing vast datasets of soil properties. In [5] the author applied Random Forest and Support Vector Machines (SVM) for soil classification based on moisture, pH, and macronutrient levels. Their model achieved an accuracy of 90.2% in predicting soil fertility. In [6] the author employed deep learning models to analyze soil images for texture classification and nutrient estimation. Their model outperformed traditional regression-based approaches, achieving an accuracy of 92% [7].

In [8] the author predicts the crop yield on the basis of location, he uses two algorithms random forest with an accuracy of 86.35% and support vector machine with an accuracy of 99.47% and crop disease detection using image processing by which a user can get the certain pesticides from the market according to the disease. In [9] the author uses both machine learning (SVM) and deep learning (LSTM, RNN) for prediction of crop with 97% accuracy and they predict the best possible cheaper crop for good yield and profit. The author in [10] uses data mining techniques to predict the crops on the basis of the climate input parameter with an accuracy of over 75%. The paper [11] is predicting the price of the crops and the prices for the next 12 months are also proposed by using machine learning and model which is implemented is Decision Tree Regressor they trained the model on several Kharif and Rabi crops for price predicting.

In [12] the Agro-Ecological (AE) zone is critical in determining which crops are appropriate for cultivation in specific areas of land. However, changing environmental and climate conditions have made it more difficult for farmers to select the appropriate crops, resulting in both time and financial losses. Traditional research in this field has struggled with issues such as poor predictions, crop variability, diversification, and a high error rate. This study introduces the Inter-fused Machine Learning with Advanced Stacking Ensemble model (IML-ASE) as a method for improving crop prediction accuracy using AE zone data. The process entails gathering data from crop recommendation datasets, preprocessing it, and then feeding it into the proposed model. The advanced stacking model is made up of several layers, with the first layer acting as a base learner using various ensemble techniques, the second layer as a meta-learner, and the third layer as a fine learner. This study addresses the challenges of modern agriculture by providing farmers with more accurate crop predictions based on AE zone characteristics. Prediction performance is measured using metrics such as mean absolute error, mean square error, root mean square error, and accuracy of 97.1%, which allows for

comparisons between predictions. This advancement aims to help farmers make informed decisions and adapt to an ever-changing agricultural landscape.

3. PRE PROCESSING

Preprocessing is a process of preparing a dataset from raw dataset. Preprocessing describes any type of processing performed on raw data in order to prepare it for further processing. The real time Manur block soil dataset is preprocessed into suitable data. This handles missing values, corrects errors in format, and resolves incorrect data. It includes data cleaning, integration, transformation, and reduction. It also adds data to missing ones while removing unwanted data that introduces noise for further analysis.

The real-life information involves inaccurate, noisy, and inconsistent data. inaccurate data that missing data can be caused by a variety of factors, including data not being collected continuously, data entry errors, technical issues with biometrics, and many others, all of which necessitate proper data preparation, noisy data can be caused by a technological issue with the data collection device, a human error during data entry, or a variety of other factors and unreliable data which is caused by factors such as data duplication, human data entry, errors in codes or names and many others, all of which necessitate data preparation and analysis.

4. FEATURE SELECTION

Feature selection is an important step in machine learning that involves selecting a subset of relevant features from the original feature set in order to reduce the feature space and improve the model's performance by lowering computational power. It's an important step in machine learning, especially when working with high-dimensional data. In real-world machine learning tasks, not all features in the dataset have an equal impact on model performance. Some features may be unnecessary, irrelevant, or even noisy. Feature selection helps to remove these, improving the model's accuracy over random guessing based on all features and increasing interpretability.

4.1 RECURSIVE FEATURE ELIMINATION

Recursive Feature Elimination, or RFE for short, is a feature selection algorithm. RFE is an effective method for removing features from a training dataset prior to feature selection. A machine learning dataset for classification or regression is comprised of rows and columns, like an excel spreadsheet. Rows are often referred to as samples and columns are referred to as features, e.g. features of an observation in a problem domain. RFE is a feature selection algorithm that uses wrappers. This means that a different machine learning algorithm is provided and used at the core of the method, which is then wrapped in RFE and used to aid in feature selection.

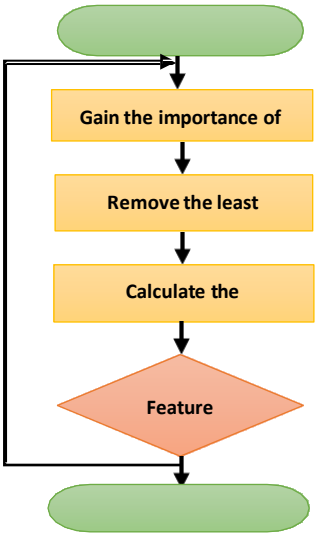

Fig. 2 Procedure of RFE

This is in contrast to filter-based feature selection, which scores each feature and selects the ones with the highest (or lowest) score. RFE can improve model performance by removing noisy or irrelevant features. Reducing the variance in predictions. Making models easier to understand. RFE train the model using the entire dataset. Rank features according to their importance scores. Remove the least significant feature(s) and retrain the model. Repeat until a stopping criterion is met, such as achieving the desired number of features or experiencing a performance drop. Fig. 2 shows the procedure of RFE of this work.

4.2 BORUTA

The primary contribution of this work is to identify the best feature selection technique, combined with a classification method, to predict the most suitable crop for cultivation based on factors such as soil and environment. Boruta is not a standalone algorithm. It sits on top of the Random Forest (RF) algorithm. In fact, the name Boruta is derived from Slavic mythology's name for the forest spirit. Random Forest is based on bagging, which is the process of creating many random samples from the training set and training a different statistical model for each one. A classification task returns the majority of model votes, whereas a regression task returns the average of the various models. Fig. 3 shows the procedure for Boruta.

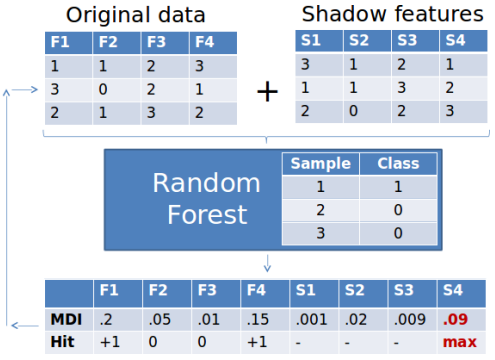

Fig. 3 Procedure for Boruta

5. INTEGRATING STACKING ENSEMBLE LEARNING

The Integrating Stacking Ensemble Learning (ISEL) model created with five base learners, namely k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Tree (DT) algorithms and with one meta classifier SVM to predict and recommend suitable crops to farmers. Before running the classifier algorithm, the optimized dataset is divided into training and testing datasets. The classifier algorithm is trained on the training dataset, and the trained classifier is used during the testing phase. The obtained results are used to forecast the crop for a real-time Manur block dataset. The classifiers used in this work are described in the subsections that follow.

5.1 K-NEAREST NEIGHBOR

K-nearest neighbors (KNN) is a supervised learning algorithm used for regression and classification. KNN attempts to predict the correct class for the test data by computing the distance between the test data and all of the training points. Then choose the K number of points that are closest to the test results. The KNN algorithm calculates the probability that the test data belongs to the classes of 'K' training data, and the class with the highest probability is chosen. In the case of regression, the value is the average of the 'K' selected training points.

5.2 SUPPORT VECTOR MACHINE

Support Vector Machines (SVMs) are supervised learning algorithms that can be applied to classification and regression tasks. The main idea behind SVMs is to find a hyperplane that best separates the various classes in the training data. This is accomplished by determining the hyperplane with the largest margin, which is defined as the distance between the hyperplane and the nearest data points from each class. Once the hyperplane has been determined, new data can be classified based on which side of the hyperplane it falls. SVMs are especially useful when the data contains a large number of features and/or has a clear margin of separation.

5.3 NAÏVE BAYES

Naive Bayes is a classification technique that is based on Bayes' Theorem with an assumption that all the features that predicts the target value are independent of each other. It computes the probabilities for each class and then selects the one with the highest probability. It has been used successfully for a variety of purposes, but it is especially effective with natural language processing (NLP) problems. Bayes' Theorem describes the probability of an event based on prior knowledge of conditions that may be relevant to that event.

5.4 RANDOM FOREST

Random forest is a machine learning technique that employs numerous individual decision trees. During the tree-building process, the optimal split for each node is determined using a set of randomly selected candidate variables. Aside from predicting outcomes in classification

and regression analyses, Random Forest can also be used to select essential variables and groups, allowing for a more in-depth understanding of variable relationships.

5.5 DECISION TREE

A decision tree is a graphical representation of various problem-solving options and how different factors are related. It has a hierarchical tree structure that begins with one main question at the top called a node and then branches out into different possible outcomes, such as: Root Node is the starting point for the entire dataset. Branches are the lines that connect the nodes. It depicts the progression from one decision to another. Internal nodes are points at which decisions are made based on input features. Leaf Nodes are the terminal nodes at the end of branches that represent final results or predictions.

6. PERFORMANCE MEASURES

In this work the following performance measures are calculated by both the 1D-CNN techniques. A set of evaluation parameters used in this work are

- Accuracy
- Precision
- Recall
- F-score

7. EXPERIMENTAL RESULTS

7.1 DATASETS

This dataset were collected from Manur soil block from the place of Manur. Manur or Manoor is a village in Tirunelveli Taluka, Tirunelveli District, Tamil Nadu, India. It is located 15 kilometers from Tirunelveli, which is both the district and sub-district headquarters for Manur village. The dataset contains block name, village name, lab no., texture, L.S, pH, E.C., O.C%, N, P, K, S, Zn, B, Fe, Mn and Cu which are the important attributes used in this work. The numbers of instances used are 943. The attribute characteristic includes nominal and numeric values. A total of 943 instances were used in which 700 were used for training and 243 were used for testing.

7.2 ENVIRONMENTAL SETUP

A set of experiments have been conducted to examine the effectiveness of machine learning algorithms in terms of crop prediction and recommendation. The simulation environment for the proposed Crop Yield Prediction System is implemented using Python, leveraging libraries such as Pandas and NumPy, scikit-learn and PyTorch.

7.3 EVALUATION OF RECURSIVE FEATURE ELIMINATION (RFE)

Recursive Feature Elimination (RFE) is a greedy optimization technique that reduces the number of input features by fitting a model repeatedly and removing the weakest feature(s) until the desired number of features is achieved. Train a model using the entire set of features. Rank features by importance (e.g., weights, coefficients). Eliminate the least important

feature(s). Repeat the process for the reduced feature set. The final result is a ranking of features and the subset with the best predictive performance.

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. 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. Fig. 4 shows the feature selection using RFE.

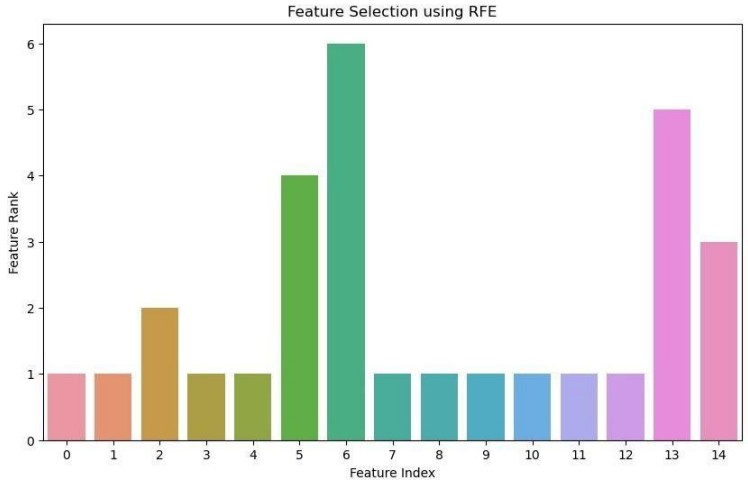

Fig. 4 Feature Selection using RFE

7.4 Evaluation of Boruta

This algorithm makes a copy of the training set features and merges them with the original ones. Creates random permutations of these synthetic features to remove any correlation between them and the target variable y; essentially, these synthetic features are randomized combinations of the original feature from which they derive. Each iteration involves a randomization of synthetic features. At each new iteration, the z-score of all original and synthetic features is calculated. A feature is considered relevant if its importance exceeds the combined importance of all synthetic features. Apply a statistical test to all original features and keep track of the results. The null hypothesis states that the importance of a feature equals the maximum importance of synthetic features. The statistical test evaluates the equality of the original and synthetic features. The null hypothesis is rejected when a feature's importance is significantly greater or less than that of synthetic features. Removes unimportant features from the original and synthetic datasets. Repeat the steps for n number of iterations until all features are removed or deemed important. Fig. 5 shows the feature selection using Boruta.

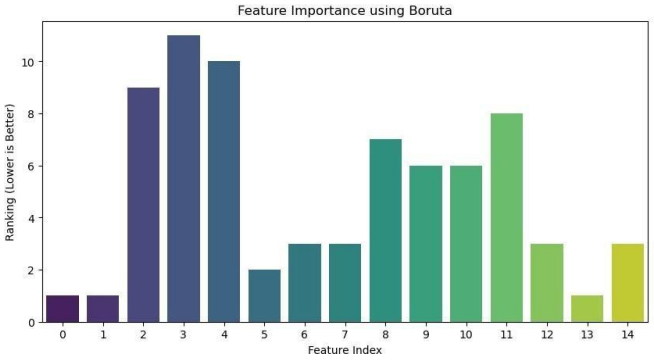

Fig. 5 Feature Selection using Boruta

7.3 Evaluation using Integrating Stacking Ensemble Learning

The Integrating Stacking Ensemble Learning (ISEL) model evaluated with five base learners, namely k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Tree (DT) algorithms and with one meta classifier SVM is highly promising in the context of this research. This comprehensive approach improves the accuracy and robustness of soil classification compared to individual algorithms or models alone. Finding the optimal settings for each algorithm and model that are time-consuming and requires extensive experimentation to achieve the best performance. Another challenge is the integration and fusion of the outputs from different classifiers and models within the ensemble architecture. Combining the predictions from multiple sources and ensuring their compatibility and synergy is a complex task that required careful handling to prevent biases or inconsistencies. Additionally the ensemble architecture required a significant amount of computational resources and processing power due to the combination of multiple algorithms and models. Training and evaluating the ensemble architecture are computationally intensive especially when dealing with large datasets. Fig. 6 shows the performance of the proposed work and the Table 5.1 shows the overall performance of the proposed work.

Fig. 6 Performance Comparison with Proposed work

Table 5 Overall Performance of the Proposed work

Feature Selection	Models	Precision (in %)	Recall (in %)	Fscore (in %)
Recursive Feature Elimination	**KNN**	80	91	85
	SVM	92	65	76
	NB	88	75	81
	RF	87	86	86
	DT	88	75	81
	Stacking	88	88	89
Boruta	**KNN**	83	91	87
	SVM	77	89	83
	NB	80	90	85
	RF	93	88	90
	DT	80	91	85
	Stacking	97	89	93

8. CONCLUSION

From the proposed work it is found that feature selection techniques namely RFE and Boruta are used to eliminate the unneeded features. Integrating Stacking Ensemble Learning technique consists of k-Nearest Neighbor (KNN), Support Vector Machine (SVM), Naïve Bayes (NB), Random Forest (RF) and Decision Tree (DT) algorithms to classify the classes. The Meta classifier SVM is used to predict and recommend the suitable crops to the farmers. The proposed work yields the accuracy of 98 %, precision of 98%, recall of 98 % and fscore of 97 % in predicting the crops.

REFERENCES

1. Chetan Raju, Ashoka D.V., Ajay Prakash B.V., (2024), "CropCast: Harvesting the Future with Interfused Machine Learning and Advanced Stacking Ensemble for Precise Crop Prediction." *Kuwait Journal of Science*, Volume 51, Issue 1, 100160, ISSN 2307-4108, https://doi.org/10.1016/j.kjs.2023.11.009.
2. Suruliandi, A., Mariammal, G., & Raja, S. P. (2021). "Crop prediction based on soil and environmental characteristics using feature selection techniques." *Mathematical and Computer Modelling of Dynamical Systems, 27*(1), 117–140. https://doi.org/10.1080/13873954.2021.1882505
3. Gunasekaran, H., Kanmani, D., & Krishnamoorthi, R. (2024). "Optimized Ensemble Learning for Enhanced Crop Recommendations: Leveraging ML for Smarter Agricultural Decision-Making.", *Engineering Proceedings, 82*(1), 95. https://doi.org/10.3390/ecsa-11-20366
4. Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). "Crop yield prediction using machine learning: A systematic literature review." *Computers and Electronics in Agriculture*, 177, 105709.
5. Dey B, Ferdous J, Ahmed R., (2024), "Machine Learning based Recommendation of Agricultural and Horticultural Crop Farming in India under the Regime of NPK.", soil pH

and Three Climatic Variables". *Heliyon*. Jan 26;10(3):e25112. doi: 10.1016/j.heliyon.2024.e25112. PMID: 38322954; PMCID: PMC10844259.

6. Gupta, R., Sharma, P., & Kumar, V. (2021). "Machine learning approaches for soil nutrient prediction. Computers and Electronics in Agriculture.", 183, 105995.
7. Zhou, T., Wu, L., & Wang, Y. (2020). "Deep Learning Techniques for Soil Classification and Nutrient Estimation.", *AI in Agriculture*, 7(2), 22-39.
8. Bondre, D. A. and S. Mahagaonkar, (2019), "Prediction of crop yield and fertilizer recommendation using machine learning algorithms," *International Journal of Engineering Applied Sciences and Technology*, vol. 04, no. 05, pp. 371–376.
9. Agarwal S. and Tarar S., (2021), "A hybrid approach for crop yield prediction using machine learning and deep learning algorithms.", *Journal of Physics: Conference Series*, vol. 1714, no. 1, p. 012012.
10. Veenadhari S., Misra B., and Singh C. D., (2014) "Machine Learning Approach for forecasting crop yield based on climatic parameters," *2014 International Conference on Computer Communication and Informatics*.
11. R. Dhanapal, A. Ajan Raj, S. Balavinayagapragathish, and J. Balaji, "Crop price prediction using supervised machine learning algorithms," *Journal of Physics: Conference Series*, vol. 1916, no. 1, p. 012042, 2021.
12. Chetan Raju, Ashoka D.V., Ajay Prakash B.V., (2024), "CropCast: Harvesting the Future with Interfused Machine Learning and Advanced Stacking Ensemble for Precise Crop Prediction." *Kuwait Journal of Science*, Volume 51, Issue 1, 100160, ISSN 2307-4108, https://doi.org/10.1016/j.kjs.2023.11.009.