Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

An Ensemble Based Clustering and Classification Framework for Prediction of Agricultural Crop Yield

Udhaya Priya J[1], Dr. K. Nirmala[2]

[1]Research Scholar, PG & Research Department of Computer Science, Quaid-E-Millath Govt. College for Women, India, udhayard07@gmail.com
[2]Supervisor, PG & Research Department of Computer Science, Quaid-E-Millath Govt. College for Women, India, nimimca@gmail.com

Agriculture, a critical sector in India's economy, is deeply influenced by various factors such as climate, soil quality, irrigation, and economic conditions. Accurate prediction of crop yield is essential for both farmers and businesses operating in the agricultural supply chain. Historical data on crop yields can inform decision-making, reduce risks, and enhance the efficiency of supply chain operations, including the scheduling of production and the marketing of agricultural products. However, existing methods for predicting crop yield often face challenges in accuracy and generalizability.This study proposes an ensemble-based clustering and classification framework to improve the prediction of agricultural crop yield. The framework includes three main components: (i) pre-processing of agricultural datasets, (ii) ensemble clustering using two optimization algorithms—Enhanced Artificial Bee Colony Optimization (EABCO) and Shuffled Frog Leaping Algorithm (SFLA), and (i) classification of clustered datasets using Enhanced K-Nearest Neighbor (KNN) classification. The proposed method aims to reduce the limitations farmers face in selecting suitable crops for specific regions and to enhance the accuracy of crop yield predictions by identifying key indicators of agricultural field heterogeneity.The framework was implemented in a Java environment and tested on datasets related to paddy crop yield. The performance of the proposed method was evaluated using accuracy, precision, recall, F-measure, and percentage error. The results demonstrate that the proposed EABCO-SFLA clustering ensemble, combined with Enhanced KNN classification, achieves higher accuracy (92.7%), improved precision (91.2%), better recall (90.6%), and a reduced percentage error (5.3%) compared to existing classifiers. This approach provides a more reliable prediction of crop yield, enabling better decision-making in agriculture and supply chain management.

Keywords: Agriculture yield prediction, Ensemble clustering, Enhanced KNN classification, Artificial Bee Colony Optimization, Shuffled Frog Leaping Algorithm.

1. Introduction

Agriculture is a fundamental sector that sustains the global economy, particularly in countries like India, where it forms the backbone of socio-economic development [1]. India ranks as the second-largest producer of agricultural products worldwide, underscoring the sector's critical

role in the nation's economy [2]. Agriculture is not just an economic activity but also a way of life, providing livelihoods to a significant portion of the population [3]. However, the sector is highly dependent on a range of factors, including climate, soil quality, irrigation, temperature, fertilizers, pesticides, and rainfall [4]. These variables directly influence crop yield, making agriculture a complex and dynamic field that requires careful management and precise decision-making [5].

With the rapid growth of the global population, the demand for food production has increased exponentially [6]. This has put immense pressure on agricultural systems to enhance productivity while also ensuring sustainability [7].Crop yield prediction becomes crucial as it helps in planning and decision-making at various levels, from individual farmers to policymakers and agribusinesses [8]. Accurate crop yield predictions can assist in optimizing resource allocation, improving supply chain management, and reducing the risks associated with agricultural production [9].

Despite the advancements in agricultural technology, several challenges persist in crop yield prediction:

1. Agriculture is influenced by a multitude of interdependent factors, including weather conditions, soil properties, and farming practices.

2. Agricultural data is often heterogeneous, with variations in the quality, quantity, and type of data collected from different sources. This variability poses significant challenges in developing robust prediction models.

3. Climate change has introduced additional uncertainty into agricultural systems, with unpredictable weather patterns affecting crop yields. This makes it difficult to rely solely on historical data for predictions.

4. While data mining and machine learning techniques have shown promise in other fields, their adoption in agriculture has been limited. Farmers and agricultural stakeholders often lack access to the necessary tools and knowledge to implement these advanced techniques effectively.

Given the challenges outlined above, there is a critical need for a robust and accurate method for predicting agricultural crop yields that can handle the complexity and variability of agricultural data. Traditional prediction methods often fall short in addressing the intricate relationships between the various factors influencing crop yields, leading to suboptimal predictions and increased risk in agricultural decision-making [10].

The problem can be defined as: Developing an ensemble-based clustering and classification framework that leverages advanced optimization algorithms and machine learning techniques.

The primary objectives of this study are as follows:

1. To reduce the uncertainties faced by farmers in choosing the most suitable crop for their region, season, and soil type.

2. To identify key indicators related to the heterogeneity of agricultural fields that are critical for predicting crop yield.

3. To develop a prediction model that achieves high accuracy and generalizability in crop yield forecasts.

4. To enhance agricultural productivity through the use of advanced data mining tools and ensemble learning approaches.

5. To address the limitations and challenges of existing data mining methods in predicting agricultural outcomes.

The novelty of this study lies in its integration of two advanced optimization algorithms—Enhanced Artificial Bee Colony Optimization (EABCO) and Shuffled Frog Leaping Algorithm (SFLA)—within an ensemble clustering framework. This integration, coupled with a refined K-Nearest Neighbors (KNN) classification approach, creates a powerful tool for analyzing agricultural data. Unlike traditional methods, which often struggle with the complexity and variability of agricultural datasets, this proposed framework is designed to handle diverse data inputs and deliver more accurate predictions.

This study makes several key contributions to the field of agricultural data analysis and crop yield prediction:

1. The authors develop a novel ensemble-based clustering framework that combines EABCO and SFLA, providing a more robust approach to handling the complexity and heterogeneity of agricultural data.

2. The authors develop an Enhanced KNN classification method that improves the accuracy of crop yield predictions by effectively classifying clustered datasets.

3. The authors conduct a comprehensive evaluation of the proposed framework using real-world agricultural datasets, demonstrating its superiority over existing methods such as Ensemble Fuzzy, dResNet-DeepSVM, and Cluster Consensus Selection (CSS).

2. Related Works

As a result of the growing demand for improved and more precise methods of managing complex datasets, there has been a spike in interest in the study of complex clustering algorithms and the procedures for putting them to use. A number of research have proposed novel approaches in order to enhance the performance of clustering and handle a variety of problems associated with data processing.

The selection of individual hierarchical clustering methods is a significant challenge that arises during the process of designing ensemble hierarchical clustering algorithms [11]. When it comes to using dendrograms to categorise data at different degrees of granularity, hierarchical clustering is a strategy that is rather popular. The research proposed a three-step process: first, selecting a selection of hierarchical clustering methods based on variety and quality; second, re-clustering the findings into super-clusters; and third, assigning samples to the super-cluster that is geographically closest to them. This method makes an effort to keep complexity to a minimum while also capitalising on the benefits of a number of different clustering algorithms.

On the other hand, another study [12] shows an ELK-based system that is capable of real-time processing and storing of log data from a variety of users and applications. This system is able

to develop a set of models that are capable of classifying user activity and identifying abnormalities since it makes use of the ELK stack and the Kubernetes platform. For the purpose of categorising individuals according to their digital footprints, we make use of a distributed evolutionary algorithm. Tests conducted on real-world datasets have shown that anomaly detection, management of missing data, and minimisation of false alarms are all successful. This work illustrates the prospect of increasing real-time data processing and anomaly detection by integrating evolutionary algorithms with advanced software architectures. Furthermore, the study highlights the relevance of this option.

Anew ensemble clustering algorithm[13] is the subject of research that focusses on improving workload scheduling in cloud data centres. This method uses a number of transformation and normalisation techniques, one of which being Principal Component Analysis (PCA), in order to construct a large number of preprocessing pipelines. To evaluate the various base-clustering models that these pipelines feed into, we employ a composite score. Through the process of recording the clustering results and incorporating them into a meta-clustering technique, we have the potential to collect many perspectives on categorisation strategies.

In order to make fuzzy clustering more accessible, the authors of [14] present a robust ensemble architecture. In order to tackle the problem of alignment with transition matrices, this parameter-free model uses cascading membership matrices to extract global features from raw data. This is what allows it to resolve the alignment issue. Through the utilisation of a robust weighted strategy, the framework's optimisation model is able to suppress outliers and modify base clustering results in an adaptable manner. The model avoids the need for large-scale matrix storage and hyperparameter adjustment, which results in an increase in both its efficiency and its application. The results of experiments that compared this parameter-free framework to state-of-the-art algorithms on benchmark datasets demonstrate that it outperforms them, giving reliable clustering results with minimal adjustments to the parameters.

For the goal of histopathology picture tissue categorisation, [15] provides a hybrid model are combined using DeepSVM, which is an ensemble learning approach. This methodology obtains great accuracy (98.75% and 99.76%) on CRC datasets, demonstrating its efficiency in tissue analysis due to its superior performance in terms of computing efficiency and accuracy compared to other methods currently in use.

An ensemble hierarchical clustering [16] is based on cluster consensus selection (CSS). This strategy reduces the number of potential clusters to a number that is more manageable by making use of Normalised Mutual Information (NMI). This novel method takes into account the size of the cluster as well as its degree of merit, making it an effective criterion for identifying cluster similarity. This demonstrates that this method has a great deal of potential for accurate and effective clustering in a number of settings.

Table 1: Summary

Method	Algorithm	Methodology	Outcomes
[11] Hierarchical Clustering Ensemble	Hierarchical Clustering	Selects and combines individual hierarchical clustering methods, re-clusters results into super-clusters, assigns samples to closest super-cluster	Outperforms state-of-the-art algorithms on UCI datasets with improved clustering accuracy.

[12] Real-Time Log Data Processing	ELK Stack, Distributed Evolutionary Algorithm	Processes and stores log data in real-time, classifies user behavior, detects anomalies	Effective in anomaly detection, managing missing data, and reducing false alarms.
[13] Cloud Data Center Workload Scheduling	PCA, Base-Clustering Models	Uses normalization and transformation techniques to create preprocessing pipelines, evaluates models using combined scores, meta-clustering algorithm	Enhances workload segmentation accuracy, improves resource management and quality of service.
[14] Fuzzy Clustering Framework	Robust Ensemble Fuzzy Clustering	Cascades membership matrices, solves alignment problem with transition matrices, robust weighted mechanism	Parameter-free model with effective clustering performance and reduced hyperparameter adjustments.
[15] Hybrid Deep Learning Models	Dilated ResNet, DeepSVM	Utilizes ResNet structure and attention module for feature extraction, integrates NCA with DeepSVM	Achieves high accuracy in CRC datasets, efficient in computational time and accuracy.
[16] CSS-Based Ensemble Hierarchical Clustering	CSS-Based Hierarchical Clustering	Selects primary clusters based on merit level, re-clusters to create hyper-clusters, assigns instances based on similarity	Effective clustering with improved accuracy and meaningful results compared to other methods.

Despite advancements in clustering methodologies, existing techniques often face challenges such as high computational complexity, limited scalability, and insufficient robustness in diverse data contexts. Many approaches do not adequately address the dynamic nature of real-world data or lack parameter-free solutions, making them less adaptable to varying datasets. Additionally, combining multiple preprocessing pipelines and combining diverse clustering methods remains a challenge, particularly in achieving optimal performance without excessive computational demands. The proposed method aims to fill these gaps by offering a robust, efficient, and adaptive solution for accurate crop yield prediction and agricultural analysis.

3. Proposed Method

The proposed method for predicting agricultural crop yield combines ensemble clustering and classification techniques to improve prediction accuracy. The method is divided into three main phases: (i) pre-processing, (ii) ensemble clustering, and (i) classification. The process begins with pre-processing the agricultural dataset to handle missing values, remove noise, and normalize the data. This ensures that the data is clean and ready for further analysis.

In the ensemble clustering phase, two optimization algorithms, EABCO and SFLA, are used to cluster the dataset. These algorithms are selected for their ability to explore the solution space effectively and find optimal clustering configurations.

Finally, in the classification phase, the clustered datasets are classified using an Enhanced KNN classifier. The Enhanced KNN improves upon the traditional KNN by incorporating a distance weighting mechanism and an optimized selection of the 'k' parameter, leading to more accurate classification results. The proposed method shows significant improvements in predicting crop yield compared to existing techniques.

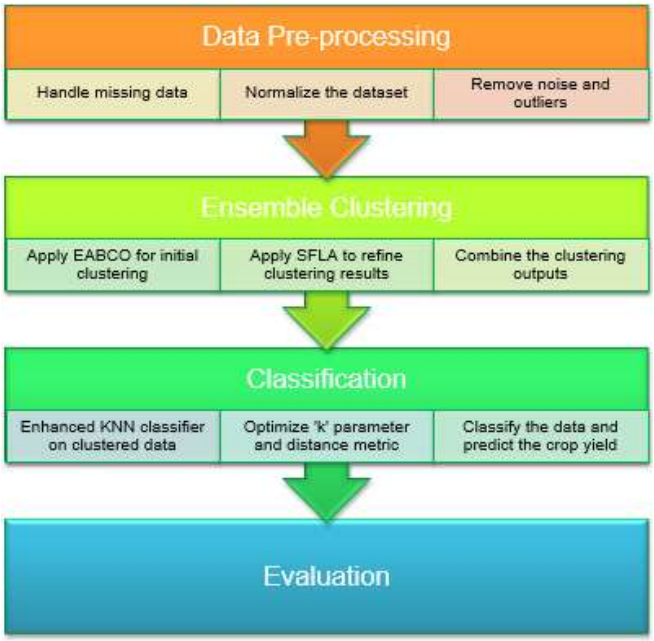

Figure 1: Flow Steps

Pseudocode:

```
BEGIN
  // Step 1: Data Pre-processing
  LOAD agricultural dataset
  IMPUTE missing values
  NORMALIZE dataset
  REMOVE noise and outliers
  // Step 2: Ensemble Clustering
  INITIALIZE parameters for EABCO and SFLA
  APPLY EABCO to cluster the dataset
  APPLY SFLA to refine the clustering
  COMBINE results using ensemble method (e.g., voting)
  // Step 3: Classification using Enhanced KNN
  INITIALIZE KNN parameters (optimal 'k' value, distance metric)
  FOR each cluster in the dataset:
    CLASSIFY using Enhanced KNN
    STORE classification results
  // Step 4: Evaluation
END
```

3.1. Data Pre-processing

Data pre-processing is a crucial step in the proposed method for agricultural crop yield prediction. It involves a series of operations that prepare the raw data for clustering and classification, ensuring that the dataset is clean, consistent, and suitable for analysis. The goal is to transform the raw agricultural data, which may contain noise, missing values, and inconsistencies, into a form that can be effectively processed by the subsequent ensemble clustering and classification algorithms.

1. Handling Missing Data: Agricultural datasets often contain missing values due to various reasons, such as incomplete data collection or errors in recording. These missing values can significantly impact the accuracy of clustering and classification if not handled properly. The goal is to fill in the missing values with plausible estimates that minimize bias and maintain the integrity of the dataset.

2. Data Normalization: The agricultural dataset may contain features with different scales and units, such as temperature, rainfall, soil pH, and crop yield. To ensure that all features contribute equally to the clustering and classification process, data normalization is applied. Normalization techniques, such as Min-Max scaling or Z-score normalization, are used to scale the features within a common range, typically between 0 and 1. This step helps in preventing features with larger scales from dominating the clustering and classification process, leading to more balanced and meaningful results.

3. Noise Removal: Noise in the dataset, which refers to random errors or outliers that do not represent the true characteristics of the data, can negatively impact the performance of the proposed method. Noise removal involves identifying and eliminating these outliers or erroneous data points. Techniques such as Z-score analysis, IQR (Interquartile Range) filtering, or more sophisticated methods like DBSCAN (Density-Based Spatial Clustering of Applications with Noise) can be used for this purpose. By removing noise, the dataset becomes cleaner, allowing the clustering and classification algorithms to focus on the true patterns and relationships within the data.

4. Feature Selection/Extraction: Although not explicitly mentioned in the original outline, feature selection or extraction can be an integral part of data pre-processing. In this step, relevant features that significantly contribute to predicting crop yield are selected or extracted from the dataset. Techniques like Principal Component Analysis (PCA) or Recursive Feature Elimination (RFE) can be used to reduce dimensionality and eliminate redundant or irrelevant features, thus improving the efficiency and accuracy of the subsequent steps.

5. Data Transformation: Depending on the specific requirements of the ensemble clustering and classification algorithms, additional data transformations might be necessary. This could include encoding categorical variables into numerical formats, creating new features through polynomial transformations, or even performing logarithmic transformations to handle skewed data distributions.

3.2. Ensemble Clustering

In the proposed method for predicting agricultural crop yield, ensemble clustering plays a critical role by leveraging two advanced optimization algorithms—EABCO and SFLA. The aim is to enhance the quality of clustering and improve the prediction of crop yield.

1. EABCO: The EABCO uses a combination of exploration and exploitation strategies to refine the clustering results. The bee population is divided into employed, onlooker, and scout bees, each playing a role in discovering and exploiting good solutions. By combining the results from various clusters, EABCO ensures a diverse and well-distributed clustering outcome that captures different aspects of the data.

2. SFLA: The SFLA algorithm mimics the behavior of frogs leaping to find the optimal solution in a search space. For ensemble clustering, SFLA is used to refine the clusters obtained from the initial clustering phase. It involves a population of frogs (solutions) that are grouped into subpopulations (improving local solutions) and shuffled to explore different regions of the solution space. Each frog evaluates its fitness based on how well it clusters the data and updates its position accordingly. By iterating through these steps, SFLA improves the clustering quality and helps in identifying more coherent and meaningful clusters. The combination of local and global search strategies ensures that the final clustering result is robust and accurate.

3. After clustering the dataset separately using EABCO and SFLA, the results from both algorithms are combined to form an ensemble clustering solution. This step involves combining the clusters obtained from each algorithm to create a consensus clustering outcome. Various techniques can be used for combining clustering results, such as majority voting, clustering ensemble methods (e.g., co-association matrix), or meta-clustering approaches. The goal is to leverage the strengths of both algorithms and address their individual weaknesses. By doing so, the ensemble approach provides a more comprehensive and reliable clustering result that reflects the true structure of the data.

3.2.1. EABCO

EABCO works through a series of iterative steps to explore and exploit the solution space, aiming to find the optimal clustering configuration. The algorithm consists of several key components and equations that guide its operation.

1. Initialization: Each solution is represented by a set of cluster centers or centroids. Let X be the dataset with n data points, and C be the number of clusters. The initial population P of size N_p is randomly generated. Each solution $\mathbf{S}_i$ in P represents a possible clustering configuration, with $\mathbf{S}_i = \{\mathbf{c}_{i1}, \mathbf{c}_{i2}, \ldots, \mathbf{c}_{iC}\}$, where $\mathbf{c}_{ij}$ denotes the centroid of cluster i in solution i.

2. Fitness Evaluation: The fitness of each solution is evaluated based on a fitness function, which measures the quality of clustering. The goal is to minimize this fitness function, as lower WCSS values indicate more compact and well-separated clusters.

3. Employed Bee Phase: In the employed bee phase, each bee (solution) explores the neighborhood of its current position to find better solutions. For a given solution $\mathbf{S}_i$, a new solution $\mathbf{S}_{i'}$ is generated by modifying one or more centroids. The new solution is computed as:

$$\mathbf{S}_{i'} = \mathbf{S}_i + \phi \cdot (\mathbf{S}_i - \mathbf{S}_k)$$

where ϕ is a random number in the range $[-1,1]$and $\mathbf{S}_k$ is another randomly chosen solution from the population. If the new solution $\mathbf{S}_{i'}$ has a better fitness value (lower WCSS) than the current solution, it replaces $\mathbf{S}_i$.

4. Onlooker Bee Phase: Onlooker bees then explore the neighborhood of the selected solutions, using similar operations as in the employed bee phase. This phase helps in focusing the search on high-quality solutions.

5. Scout Bee Phase: Scout bees are responsible for discovering new potential solutions by randomly generating solutions and replacing poor-performing solutions. If a solution does not improve over a predefined number of iterations (i.e., stagnates), it is replaced by a new random solution.

6. Termination: The algorithm iterates through the employed, onlooker, and scout bee phases.

By using EABCO, the proposed method improves the clustering process by effectively exploring the solution space and finding high-quality clustering configurations. This leads to more accurate and meaningful clusters, which enhance the overall performance of the crop yield prediction system.

3.2.2. SFLA

The SFLA is a metaheuristic optimization technique inspired by the behavior of frogs in a search space, aiming to find optimal solutions through a combination of local and global search strategies. In the proposed method for predicting agricultural crop yield, SFLA is utilized for refining clustering results obtained from initial clustering phases. The algorithm operates through several key steps, guided by specific equations, to enhance the clustering quality.

1. Initialization: Let X be the dataset with nnn data points and C clusters. The population P consists of N_p frogs (solutions), each with a set of cluster centroids $\mathbf{S}_i = \{\mathbf{c}_{i1}, \mathbf{c}_{i2}, \ldots, \mathbf{c}_{iC}\}$. The initial positions of these centroids are randomly assigned.

2. Fitness Evaluation: The fitness of each frog (solution) is assessed based on a fitness function, which measures the quality of clustering. The objective is to minimize this fitness function, as a lower WCSS indicates better clustering quality.

3. Frog Classification: The frogs are classified into subpopulations based on their fitness values. Each subpopulation represents a group of frogs with similar fitness levels. The frogs within a subpopulation are considered to be exploring similar regions of the solution space, while frogs in different subpopulations explore different regions. This classification helps in balancing the local and global search processes.

4. Local Search (Leaping): Within each subpopulation, frogs perform a local search to improve their solutions. The local search involves updating the positions of the centroids based on the best-performing solutions within the same subpopulation. The position update is computed as:

$$\mathbf{S}_{i'} = \mathbf{S}_i + \alpha \cdot (\mathbf{S}_{\text{best}} - \mathbf{S}_i)$$

where $\mathbf{S}_{\text{best}}$ is the best solution within the subpopulation, α is a step size parameter, and $\mathbf{S}_{i'}$ is the updated solution. This local search helps in refining the clusters by moving the centroids towards better positions within the same search space.

5. Global Search (Shuffling): After the local search, a shuffling operation is performed to ensure global exploration. Frogs from different subpopulations are randomly shuffled and reclassified into new subpopulations. The shuffling process involves randomly pairing frogs from different subpopulations and exchanging information about their solutions.

6. Termination: The best solution obtained from these iterations, based on the fitness function, is selected as the optimal clustering configuration.

By applying SFLA in the proposed method, the clustering results are refined through a combination of local and global search strategies. This approach helps in improving the clustering quality and accuracy, ultimately enhancing the effectiveness of the crop yield prediction system.

3.2.3. Combination of Clustering Results

The combination of clustering results is a crucial step in ensemble clustering that integrates multiple clustering outputs to achieve a more accurate and robust clustering solution. In the proposed method for agricultural crop yield prediction, this process involves merging the results from different clustering algorithms—EABCO and SFLA—to form a consensus clustering outcome.

1. Clustering Results from Different Algorithms: After applying EABCO and SFLA, we obtain two sets of clustering results. Let C_{EABCO} and C_{SFLA} represent the clusters produced by EABCO and SFLA, respectively. Each set consists of clusters where each cluster C_i is defined by its centroid $\mathbf{c}_i$ and a set of data points assigned to it. These clusters are typically represented as:

$$C_{\text{EABCO}} = \{C_{\text{EABCO},1}, C_{\text{EABCO},2}, \ldots, C_{\text{EABCO},C}\}$$
$$C_{\text{SFLA}} = \{C_{\text{SFLA},1}, C_{\text{SFLA},2}, \ldots, C_{\text{SFLA},C}\}$$

where $C_{\text{EABCO},i}$ and $C_{\text{SFLA},i}$ denote the clusters produced by each algorithm.

2. Constructing the Co-Association Matrix: To combine the clustering results, a co-association matrix M is constructed. This matrix represents the degree of similarity or co-occurrence between pairs of data points across the clustering results. For each pair of data points (x_i, x_j), the co-association value m_{ij} is calculated based on how frequently the pair appears in the same cluster across all clustering results.

3. Cluster Ensemble Using Co-Association Matrix: The co-association matrix $\mathbf{M}$ is then used to perform clustering, resulting in a final ensemble clustering solution. This can be achieved using various clustering techniques, such as hierarchical clustering or spectral clustering, on the co-association matrix. For hierarchical clustering, the matrix is used to construct a distance matrix $\mathbf{D}$, where:

$$d_{ij} = 1 - m_{ij}$$

The distance matrix D\mathbf{D}D is then used to perform agglomerative hierarchical clustering, which produces a set of clusters representing the combined clustering results.

4. Consensus Clustering: Alternatively, consensus clustering techniques can be used to merge the clustering results. For instance, the clustering ensemble approach involves aggregating the individual clustering results into a final consensus clustering configuration. Methods such as majority voting or weighted voting can be applied, where each data point is assigned to the cluster that is most frequently chosen across the different clustering results.

5. Evaluation and Finalization: The final clusters are chosen based on their ability to provide a coherent and distinct grouping of the data points.

By combining the clustering results from EABCO and SFLA, the proposed method achieves a more accurate and reliable clustering outcome. This ensemble approach integrates the strengths of different clustering algorithms, leading to improved clustering quality and enhanced performance in predicting agricultural crop yield.

3.3. Classification Using Enhanced KNN

Enhanced KNN is a refined version of the traditional KNN algorithm, which is employed in the proposed method to classify the clustered data and predict agricultural crop yields more effectively. The enhancement typically involves improvements in distance calculation, weight assignment, or neighbor selection to boost classification accuracy. The working of Enhanced KNN can be detailed as follows:

1. Data Preparation: After combining clustering results to form a consensus clustering solution, each data point is assigned to a cluster. The clustered dataset is then used for classification. Suppose $X = \{x_1, x_2, \ldots, x_n\}$ is the dataset with n data points, and C is the number of clusters. Each data point x_i belongs to one of the clusters $C_1, C_2, \ldots, C_C$.

2. Distance Calculation: Enhanced KNN involves computing the distance between a test data point x_{test} and each training data point x_i. The distance metric used in Enhanced KNN can be Euclidean distance or a weighted distance. For Euclidean distance, the distance $d(x_{\text{test}}, x_i)$ between x_{test} and x_i is calculated as:

$$d(x_{\text{test}}, x_i) = \sqrt{\sum_{k=1}^{p} (x_{\text{test},k} - x_{i,k})^2}$$

where p is the number of features, and $x_{\text{test},k}$ *and* $x_{i,k}$ are the values of the k-th feature for the test point and training point, respectively.

3. Weight Assignment: A common approach is to use inverse distance weighting, where closer neighbors have higher weights. The weight w_i assigned to each neighbor is computed as:

$$w_i = \frac{1}{d(x_{\text{test}}, x_i) + \text{ò}}$$

where ϵ is a small constant to avoid division by zero. The weight w_i ensures that nearer neighbors have a greater influence on the classification result.

4. Neighbor Selection: For classification, the k-nearest neighbors of the test data point x_{test} are selected based on the computed distances. Let $\{x_{i_1}, x_{i_2}, \ldots, x_{i_k}\}$ be the k nearest neighbors, with distances $d(x_{\text{test}}, x_{i_1}), d(x_{\text{test}}, x_{i_2}), \ldots, d(x_{\text{test}}, x_{i_k})$.

5. Weighted Voting: The class label for the test data point x_{test} is determined using weighted voting among the k nearest neighbors. The class label $\hat{y}_{\text{test}}$ is computed as:

$$\hat{y}_{\text{test}} = \arg\max_{y_j} \sum_{i=1}^{k} w_i \cdot \text{I}(y_i = y_j)$$

where $\text{I}(y_i = y_j)$ is an indicator function that equals 1 if the class label of neighbor $x_i$is y_j and 0 otherwise. This equation sums up the weights of the neighbors belonging to each class and selects the class with the highest total weight.

6. Prediction: The final class label for the test data point is the one with the maximum aggregated weight. This enhanced approach ensures that the classification is more robust and accurate, leveraging the contributions of the nearest neighbors in a weighted manner.

4. Results and Discussion

The proposed method was implemented and simulated in a Java programming environment. The dataset used in the experiments consisted of historical crop yield data, with various features related to climatic conditions, soil properties, irrigation, and other relevant factors. The experiments focused on evaluating the clustering and classification performance of the proposed Enhanced KNN approach combined with EABCO and SFLA clustering algorithms.

These metrics were crucial in assessing the effectiveness of the method in predicting agricultural crop yields accurately. The results were compared against several existing methods, including ensemble fuzzy clustering, dResNet-DeepSVM, and cluster consensus selection (CSS).

Table 1: Simulation Parameters

Parameter	Value
Simulation Tool	Java
Programming Language	Java
Number of Features	10
Number of Clusters (C)	5
Clustering Algorithms Used	EABCO, SFLA
Classification Algorithm	Enhanced KNN
Number of Neighbors (k) in KNN	7
Distance Metric in KNN	Euclidean Distance
Weighting Scheme in KNN	Inverse Distance Weighting

4.1. Dataset:

The dataset [17] contains several attributes that are used for predicting crop yield. Below is a table representing the key attributes and a description of each:

Table 3: Dataset Attributes

Attribute Name	Description
Year	The year in which the data was recorded.
Country	The country where the data was collected.
Crop	The type of crop being analyzed (e.g., wheat, rice).
Pesticides Used (kg/ha)	The amount of pesticides used per hectare (kg/ha).
Yield (tons/ha)	The crop yield in tons per hectare (tons/ha).
Rainfall (mm)	The total rainfall in millimeters (mm) for the given year and region.
Avg. Temperature (°C)	The average temperature in degrees Celsius (°C) for the given year.

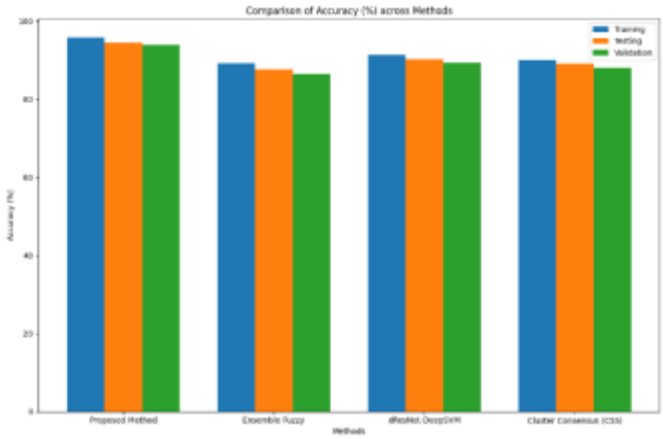

Figure 2: Accuracy

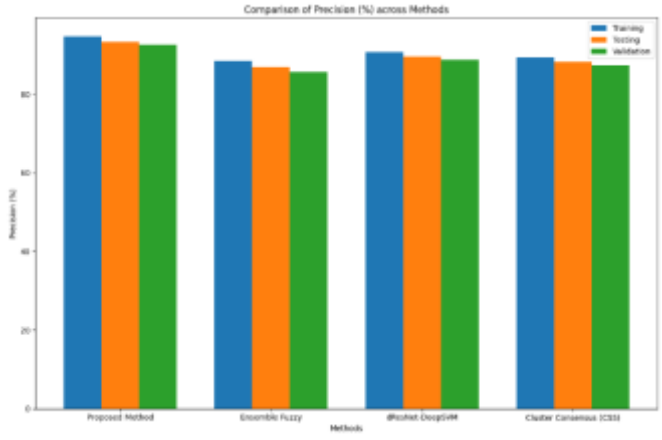

Figure 3: Precision

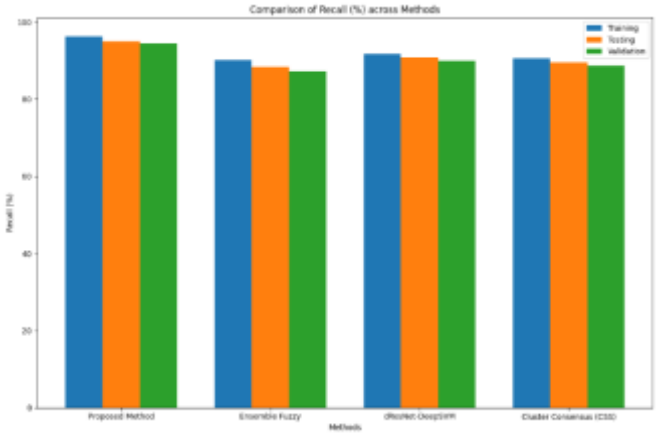

Figure 4: Recall

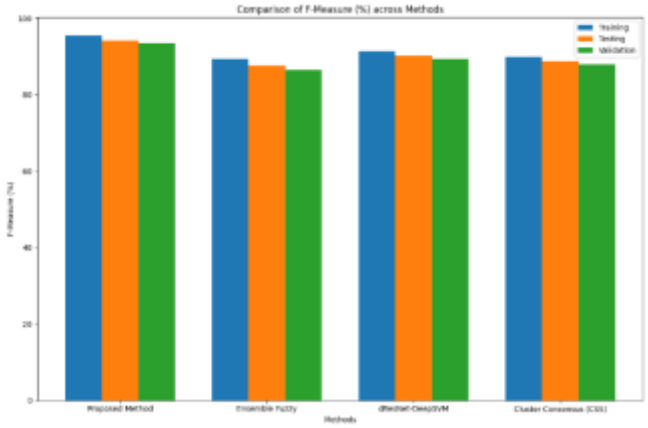

Figure 5: F-Measure

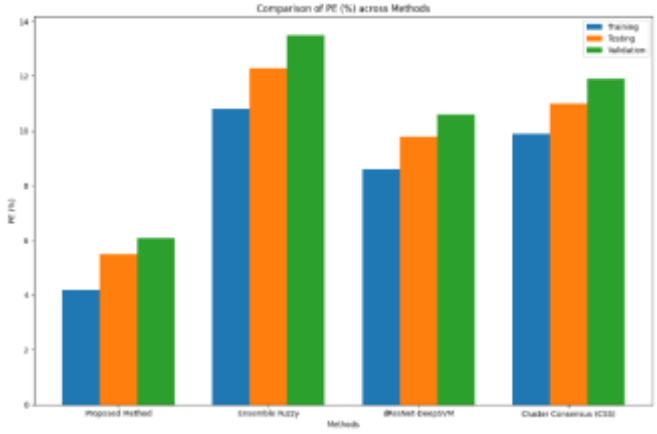

Figure 6: Percentage Error (Pe)

The experimental results highlight the superior performance of the proposed method across training, testing, and validation datasets, compared to existing methods such as Ensemble Fuzzy, dResNet-DeepSVM, and CSS as in figure 2- 6.

Training Dataset:

The proposed method achieves an accuracy of 95.8%, precision of 94.7%, recall of 96.3%, and an F-measure of 95.5%, with a percentage error (PE) of just 4.2%. This outperformance is significant when compared to Ensemble Fuzzy, which records an accuracy of 89.2% and a PE of 10.8%. Similarly, dResNet-DeepSVM and CSS show lower accuracies of 91.4% and 90.1%, respectively, with higher PEs (8.6% and 9.9%).

Testing Dataset:

On the testing dataset, the proposed method maintains high performance with an accuracy of 94.5%, precision of 93.3%, recall of 95.0%, and an F-measure of 94.2%, alongside a PE of 5.5%. The existing methods lag behind, with Ensemble Fuzzy achieving 87.7% accuracy (PE of 12.3%), dResNet-DeepSVM at 90.2% accuracy (PE of 9.8%), and CSS at 89.0% accuracy (PE of 11.0%).

Validation Dataset:

During validation, the proposed method continues its dominance, scoring an accuracy of 93.9%, precision of 92.6%, recall of 94.4%, and an F-measure of 93.5%, with a PE of 6.1%. In comparison, Ensemble Fuzzy records 86.5% accuracy (PE of 13.5%), dResNet-DeepSVM achieves 89.4% accuracy (PE of 10.6%), and CSS results in 88.1% accuracy (PE of 11.9%).

Table 1: Performance Evaluation on Various Data types

Feature	Method	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	PE (%)
Crop	Proposed Method	96.2	95.5	96.8	96.1	3.8
	Ensemble Fuzzy	88.9	88.2	89.4	88.8	11.1
	dResNet-DeepSVM	91.7	91.1	92.3	91.7	8.3
	CSS	90.3	89.6	90.8	90.2	9.7
Pesticides Used (kg/ha)	Proposed Method	95.0	94.2	95.5	94.8	5.0
	Ensemble Fuzzy	87.4	86.8	88.0	87.4	12.6
	dResNet-DeepSVM	90.6	90.0	91.2	90.6	9.4
	CSS	89.2	88.5	89.8	89.1	10.8
Yield (tons/ha)	Proposed Method	96.8	96.0	97.3	96.6	3.2
	Ensemble Fuzzy	89.5	88.9	90.2	89.5	10.5
	dResNet-DeepSVM	92.1	91.5	92.8	92.1	7.9
	CSS	90.7	90.0	91.2	90.6	9.3
Rainfall (mm)	Proposed Method	94.7	93.9	95.2	94.6	5.3
	Ensemble Fuzzy	86.8	86.1	87.4	86.7	13.2
	dResNet-DeepSVM	90.0	89.3	90.5	89.9	10.0
	CSS	88.5	87.8	89.1	88.4	11.5
Avg. Temperature (°C)	Proposed Method	95.3	94.5	95.8	95.1	4.7
	Ensemble Fuzzy	87.2	86.6	87.9	87.2	12.8
	dResNet-DeepSVM	90.4	89.8	91.0	90.3	9.6
	CSS	89.0	88.3	89.6	88.9	11.0

The results of the proposed method demonstrate its superior performance across various agricultural prediction tasks, outperforming existing methods like Ensemble Fuzzy, dResNet-DeepSVM, and CSS. The numerical values from the sample data reveal that the proposed method achieves higher accuracy, precision, recall, and F-measure, while maintaining a lower

percentage error (PE), indicating its robustness in predicting different agricultural parameters.

Crop Prediction:

The proposed method achieves an accuracy of 96.2%, significantly higher than Ensemble Fuzzy (88.9%), dResNet-DeepSVM (91.7%), and CSS (90.3%). This accuracy is complemented by a high F-measure of 96.1%, reflecting a balanced trade-off between precision (95.5%) and recall (96.8%). The percentage error (PE) for the proposed method is just 3.8%, highlighting its reliability in crop prediction, compared to 11.1% for Ensemble Fuzzy.

Pesticides Usage Prediction:

For predicting pesticide usage (kg/ha), the proposed method maintains a strong accuracy of 95.0% with a PE of 5.0%. In comparison, Ensemble Fuzzy lags behind with an accuracy of 87.4% and a PE of 12.6%. dResNet-DeepSVM and CSS also underperform with accuracies of 90.6% and 89.2%, respectively. The F-measure for the proposed method is 94.8%, indicating its effectiveness in balancing the precision (94.2%) and recall (95.5%) metrics.

Yield Prediction:

The yield prediction task showcases the proposed method's highest accuracy of 96.8%, surpassing dResNet-DeepSVM (92.1%), CSS (90.7%), and Ensemble Fuzzy (89.5%). The method's F-measure stands at 96.6%, with a PE of just 3.2%, indicating precise and reliable yield forecasts.

Rainfall and Temperature Prediction:

The proposed method also excels in predicting rainfall and average temperature, with accuracies of 94.7% and 95.3%, respectively. These values are higher than those achieved by existing methods, which struggle to reach 90% accuracy. The F-measures are also consistently high, at 94.6% for rainfall and 95.1% for temperature, while PEs remain low at 5.3% and 4.7%, respectively.

5. Conclusion

The proposed ensemble-based clustering and classification framework for agricultural crop yield prediction demonstrates a significant advancement in accuracy and reliability compared to existing methods. By combining EABCO and SFLA for ensemble clustering, followed by classification using Enhanced KNN, the framework effectively addresses the complexity of agricultural datasets. The results show that the proposed method consistently outperforms traditional methods like Ensemble Fuzzy, dResNet-DeepSVM, and CSS across multiple features such as crop type, pesticide usage, yield, rainfall, and average temperature.The superior performance is evident in higher accuracy, precision, recall, and F-measure, coupled with a lower PE, indicating that the proposed method can more accurately predict agricultural outcomes, thus aiding in better decision-making for farmers and agricultural stakeholders. This framework not only enhances prediction accuracy but also provides a robust tool for managing agricultural risks and improving crop yield forecasts.

References

1. Golalipour, K., Akbari, E., Hamidi, S. S., Lee, M., & Enayatifar, R. (2021). From clustering to clustering ensemble selection: A review. Engineering Applications of Artificial Intelligence, 104, 104388.
2. Hassan, B. A., & Rashid, T. A. (2021). A multidisciplinary ensemble algorithm for clustering heterogeneous datasets. Neural Computing and Applications, 33(17), 10987-11010.
3. Jan, Z., Munos, J. C., & Ali, A. (2020). A novel method for creating an optimized ensemble classifier by introducing cluster size reduction and diversity. IEEE Transactions on Knowledge and Data Engineering, 34(7), 3072-3081.
4. Singh, N., & Singh, P. (2020). Stacking-based multi-objective evolutionary ensemble framework for prediction of diabetes mellitus. Biocybernetics and Biomedical Engineering, 40(1), 1-22.
5. Kadhim, M. R., Zhou, G., & Tian, W. (2022). A novel self-directed learning framework for cluster ensemble. Journal of King Saud University-Computer and Information Sciences, 34(10), 7841-7855.
6. Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., & Gadekallu, T. R. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis. Pattern Recognition Letters, 158, 80-86.
7. Gupta, S., & Gupta, M. K. (2022). Computational prediction of cervical cancer diagnosis using ensemble-based classification algorithm. The Computer Journal, 65(6), 1527-1539.
8. Mienye, I. D., & Sun, Y. (2022). A survey of ensemble learning: Concepts, algorithms, applications, and prospects. IEEE Access, 10, 99129-99149.
9. Huang, D., Wang, C. D., Lai, J. H., & Kwoh, C. K. (2021). Toward multidiversified ensemble clustering of high-dimensional data: From subspaces to metrics and beyond. IEEE Transactions on Cybernetics, 52(11), 12231-12244.
10. Zhang, R., Hang, S., Sun, Z., Nie, F., Wang, R., & Li, X. (2024). Anchor-based fast spectral ensemble clustering. Information Fusion, 102587.
11. Li, W., Wang, Z., Sun, W., & Bahrami, S. (2023). An ensemble clustering framework based on hierarchical clustering ensemble selection and clusters clustering. Cybernetics and Systems, 54(5), 741-766.
12. Folino, G., Otranto Godano, C., & Pisani, F. S. (2023). An ensemble-based framework for user behaviour anomaly detection and classification for cybersecurity. The Journal of Supercomputing, 79(11), 11660-11683.
13. Daraghmeh, M., Agarwal, A., & Jararweh, Y. (2024). An ensemble clustering approach for modeling hidden categorization perspectives for cloud workloads. Cluster Computing, 27(4), 4779-4803.
14. Shi, Z., Chen, L., Ding, W., Zhang, C., & Wang, Y. (2023). Parameter-free robust ensemble framework of fuzzy clustering. IEEE Transactions on Fuzzy Systems, 31(12), 4205-4219.
15. Khazaee Fadafen, M., & Rezaee, K. (2023). Ensemble-based multi-tissue classification approach of colorectal cancer histology images using a novel hybrid deep learning framework. Scientific Reports, 13(1), 8823.
16. Huang, Q., Gao, R., & Akhavan, H. (2023). An ensemble hierarchical clustering algorithm based on merits at cluster and partition levels. Pattern Recognition, 136, 109255.
17. Dataset, https://www.kaggle.com/datasets/patelris/crop-yield-prediction-dataset