Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 10
October 2023

Enhancing Crop Yield Prediction Through Advanced Data Mining Techniques

A. Chitradevi[1], Dr. N. Tajunisha[2]
1Research Scholar & Assistant Professor, Sri Ramakrishna College of Arts and Science for Women
2Head & Professor, Sri Ramakrishna College of Arts and Science for Women

Abstract
Crop yield prediction aids resource allocation and agricultural decision-making. Integration of several data sources and effective data preprocessing and feature selection are needed for accurate prediction. Machine learning procedures improve model performance and interpretability via normalization and feature selection. This paper has proposed crop yield prediction using ensemble-based normalization and feature selection methods with SVM Classification. The ensemble normalization has utilized with Average filling, Weighted K-means clustering and Decision tree algorithms. Weighted K-means clustering and decision tree assigns values to samples based on their distances from cluster centers to show data distribution. An average filling fills missing values with the average of their properties, completing the dataset for analysis. Next, the feature selection has utilized Random Forest (RF), Logistic Regression (LR), PCA and Elastic Net selects important features. Principal components analysis optimizes representation and feature selection by selecting orthogonal components that best reflect data variation. The last step is classification using Support Vector Machine (SVM). The SVM model has classify the new instances using the important features. To improve crop yield production using rainfall, humidity, N, P, K and pH attributes are considered. These factors are crucial to crop health and growth. The SVM classification has achieved 91% accuracy while using the ensemble normalization and feature selection methods used.

Keywords: Crop Yield**,** Classification, Ensemble, Feature Selection, Normalization

I. Introduction

Agriculture, as the major source of food, is under tremendous pressure to deliver bigger yields while dealing with the unpredictability of climate change. Crop yield prediction has emerged as a critical area of study and application in this environment [1]. Researchers and agricultural specialists are attempting to create accurate and effective ways of forecasting crop yields by using the capabilities of new technologies such as machine learning, data analytics, and remote sensing [2]. These forecasts not only help farmers make educated choices, but they also help governments devise methods to reduce food shortages and stabilize economies. This study examines the relevance of agricultural production prediction, investigates the approaches used, and discusses the consequences for converting agriculture into a more resilient and productive industry. We uncover the potential of this emerging sector to revolutionize global food systems and secure a sustainable future for future generations by conducting an in-depth investigation of it [3].

The performance and interpretability of prediction models in machine learning may be greatly improved by preprocessing processes like dataset normalization and feature selection. Normalization ensures that characteristics are on analogous scales, reducing the impact of variables with greater magnitudes [4]. On the other hand, feature selection seeks to identify the most pertinent predictors by removing irrelevant or redundant features, thereby enhancing model efficiency and interpretability [5].

Normalization techniques such as standardization and min-max scaling have historically been used extensively [6]. Nonetheless, recent advancements in ensemble methods have demonstrated optimistic results in overcoming the limitations of conventional approaches [7]. Ensemble methods predict by combining multiple models or algorithms, capitalizing on the strengths of each component to improve overall performance [8]. Ensemble methods provide the potential for more robust and accurate preprocessing in normalization and feature selection [9].

This abstract describes an ensemble approach incorporating multiple normalization and feature selection techniques [10]. We propose a novel method for normalizing datasets that combines

weighted K-means clustering, average filling, and Decision Tree Regressor [11]. This combination permits adaptive normalization by allocating various weights to data points following their proximity to cluster centroids [12]. In addition, the average infill technique manages absent values, ensuring that valuable data is preserved during the normalization procedure. The Decision Tree Regressor identifies nonlinear relationships, enhancing normalization [13].

In addition, ensemble feature selection is introduced by integrating Random Forest (RF), Logistic Regression (LR), and Elastic net with Principal Component Analysis (PCA). Using the assets of multiple algorithms, this ensemble approach seeks to identify the most informative features [14]. While LR and Elastic Net use regularization techniques to select meaningful predictors, RF provides a robust feature ranking based on variable importance. PCA reduces dimensionality while preserving important features, improving model efficiency [15].

Our proposed method provides a comprehensive paradigm for dataset normalization and feature selection by integrating these ensemble methods [16-18]. The combination of weighted K-means clustering, average filling, and Decision Tree Regressor guarantees a more precise and trustworthy normalization procedure [19-23]. The ensemble feature selection approach using RF, LR, Elastic net, and PCA guarantees the retention of only the most informative features, thereby enhancing model interpretability and minimizing over fitting [24-25]. In the subsequent sections, we will delve into our proposed method for normalizing datasets and selecting features using ensemble methods [26-27]. Experimental results on various datasets will demonstrate our method's efficacy and efficiency, emphasizing its potential for enhancing predictive modelling tasks in real-world applications [28].

In the realm of agricultural analysis, the accurate assessment of crop yields hinges upon vital soil attributes such as Nitrogen content, Phosphorous content, Potassium content, rainfall and soil pH value [29]. These factors exert a profound influence on crop health, growth, and productivity, making them indispensable considerations in predictive models and decision-making processes aimed at optimizing agricultural outcomes [30].

The primary contributions and objectives of this manuscript may be summarized as follows.

- Dataset Normalization
- Ensemble feature selection
- Classification using SVM

The remainder of this paper is structured as follows. Numerous authors address a variety of crop yield prediction strategies in Section 2. The proposed model is shown in Section 3. Section 4 summarizes the results of the investigation. Section 5 concludes with a discussion of the result and future work.

1.1 MOTIVATION OF THE PAPER

This study offers a unique ensemble-based approach for crop yield dataset normalization that combines weighted K-means clustering with average filling, with the goal of properly capturing data distribution and handling missing values. Furthermore, for feature selection, the technique utilizes an ensemble of Random Forest, Logistic Regression, and Elastic Net, offering full insights into feature significance and relevance. By combining these methodologies, the suggested strategy improves data preparation quality, resulting in enhanced machine learning model performance and interpretability through a better knowledge of feature contributions and dataset properties.

Ii. Background Study

A. Lakshmanarao et al. [1] the author offer a number of models that include a mixture of traditional hand-crafted features and automatically retrieved embedding features, as well as the ensemble of analyzers that acquires knowledge from these features. The author proposes a two-dimensional taxonomy of ensembles of classifiers and features to categories these various methods. Banerjee, R. et al. [3] Using ensemble learning, the author were able to build a method for resolving disputes and effectively apply it to the problem of classifying crops. To better classify maize, soybeans, rice, and cotton in remote sensing, the author have

developed a hybrid approach that combines machine learning with symbolic reasoning. An ensemble learner, consisting of a decision tree, a neural network, and a support vector machine, represents the machine learning endeavour. Symbolic argumentation was used to settle contentious circumstances, such as when basic classifiers cannot agree on an instance.

Fayyazifar, N., &Samadiani, N. [7] the author introduces the ensemble Bagging approach to machine learning for IDS. For network-wide anomalous packet identification, the Bagging with REPTree basis classifier was suggested. The test dataset and 10-fold cross validation were used to assess the suggested approach. Accuracy, speed, and number of false positives were some of the metrics used to evaluate a classifier's effectiveness. Standard machine learning methods were used as a comparison for the method's efficacy.

J. Dan [9] the author explored several feature selection strategies for software defect prediction 3, and found that focusing on a small number of high-quality features significantly improves AUC. The author also demonstrated how effective ensemble learning was when applied to unbalanced datasets that include duplicate features. It was suggested to use a classifier that uses ensemble learning with two different variables. Across six different datasets, experimental results showed that greedy forward selection significantly outperformed correlation-based forward selection.

Kaur, I., & Kaur, A. [10] Using a classifier's success in differentiating between thriving and wilting patches of vegetation, the suggested ensemble feature selection process ranked the spectral properties of a hyperspectral dataset from most to least relevant. The best classification results were obtained using the top 15 features from the feature ranking list, and the top two features attained the same CV error as all 215 spectral features combined. Based on the results, feature selection need to be included as a primary pre-processing step in hyperspectral picture analysis. Feature selection reduces the burden of post-processing (such as intricate interpretation and high computing cost) while simultaneously enhancing classification precision.

Moghimi, A. et al. [12] to make the most of the power of ensemble learners while mitigating the drawbacks of high dimensionality, the author mix bagging and boosting with feature selection in this research. The author compared the results obtained with feature selection alone with those obtained with Select-Bagging and Select-Boost. These ensemble methods were trained on two separate sentiment datasets, one manually labelled and one automatically labelled in a semi-supervised fashion, utilizing four base learners and 10 feature subset sizes (composed of features picked via ROC).

Rai, A. [18] these authors research mapped crop distribution in the southern part of Jishan County, Shanxi Province, China. The author segmented photos using various preferred scales in order to extract the most informative features from each scale and apply them in these authors analysis. The author primarily chose some scales that can segment images well using the ESP tool and defined them as "preferred" scales because including all scales in the final multiscale weighted classification model would result in excessive computation loads and information redundancy.

Safiyari, A., & Javidan, R. [20] the biotechnology area relies heavily on machine learning methods for illness detection. The information learned via machine learning may be utilised to create expert systems that aid in the diagnosis and prognosis of many different diseases. In this study, the author explore the use of data mining strategies for forecasting skin diseases. When the gradient boosting ensemble approach was used on the RNC dataset for skin diseases, the resulting accuracy was 99.68 percent, which was higher than the accuracy achieved by the feature selection method. On the skin disease dataset, the author achieved the best accuracy reported in the literature to date.

2.1 Problem Definition

In machine learning pipelines, normalizing datasets and selecting relevant features using ensemble techniques is a common challenge this research attempts to solve. While feature selection seeks to discover the most informative and discriminatory characteristics for constructing strong models, normalization is crucial to guarantee that the data is

uniform and similar across features. This research offers a comprehensive solution to these difficulties by proposing an ensemble-based approach that combines weighted K-means clustering, average filling, RF, LR, Elastic net, PCA, and SVM. The normalization methods provide a true picture of the data distribution and deal with missing values. In contrast, the ensemble feature selection strategy uses some strategies to zero down on the most important features while decreasing the dataset's dimensionality. The ultimate goal of the work is to use these methods to improve the precision and interpretability of machine learning models across various settings.

III. Materials And Method

In this section, we provide a thorough strategy for crop yield dataset normalization, ensemble feature selection, and classification using SVM. To normalize the dataset, we employ weighted K-means clustering to capture the data distribution accurately, average filling to handle missing values and a decision tree regressor to address outliers. We utilize an ensemble approach combining RF with Logistic Regression (LR) and Elastic net for feature selection. This ensemble method allows us to obtain a comprehensive understanding of feature importance. To further enhance feature selection, PCA is applied to reduce dimensionality while preserving important characteristics. Finally, for classification, we employ SVM, a powerful algorithm that maximizes the margin between different classes, resulting in robust decision boundaries. We want to increase the performance and interpretability of our classification model by combining these strategies.

3.1 Dataset collection

The dataset has collected using https://www.kaggle.com/datasets/siddharthss/crop-recommendation-dataset. this dataset contains 8 attributes with 2200 data's. This dataset also contains various crops with temperature and humidity.

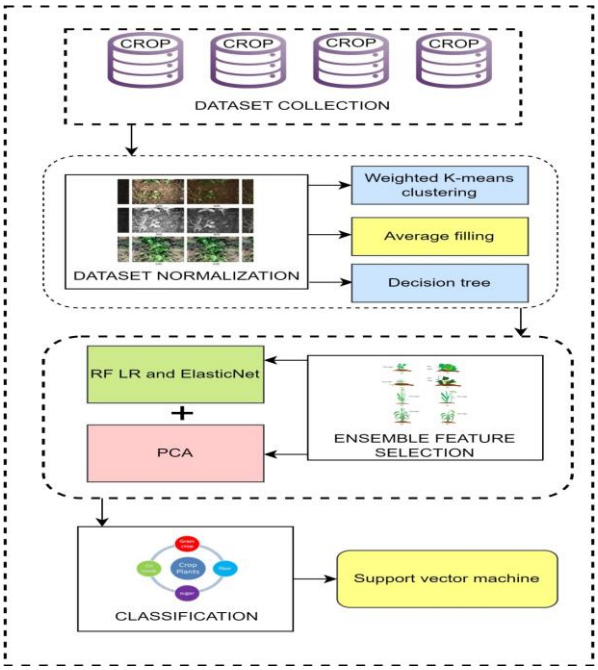

Figure 1: overall flow diagram

Figure 1 presents an overall flow diagram depicting the sequential progression of a complex process. The diagram consists of interconnected components and directional arrows, each representing a distinct stage or action within the process.

3.2 Ensemble based Dataset Normalization

After data collection, the collected dataset undergoes several normalization techniques. Weighted K-means clustering is employed to identify clusters of similar data points, with weights assigned based on their relevance. This allows the algorithm to prioritize certain data points during normalization. The average filling is applied to handle missing or incomplete data, replacing missing values with the average value of available data for each feature. This ensures the dataset remains complete and maintains statistical properties.

3.2.1 K-means clustering

K-Means, one of the most basic unsupervised learning algorithms, takes on the ever-present grouping problem. Assuming k clusters, the method provides a fast and easy way to divide data into distinct groups J. Dan (2022).

3.2.2 Weighted k-means

Weighted k-means is a variant of the classic k-means clustering technique that uses variable values for each data point at various clustering phases. The standard k-means method gives each data point the same importance score regardless of its relevance. However, in practice, certain details may be more important than others for distinguishing across clusters. Because it assigns various values to each data point based on its relevance, the weighted k-means method can resolve this problem. The weights may be learned with the help of an algorithm, or they can be set beforehand. During the clustering process, the algorithm gives more weight to the most important data and less to the less important. This is particularly beneficial in scenarios like environmental analysis, where factors such as temperature, humidity, pH value, and rainfall hold differing levels of importance, leading to more accurate and insightful cluster assignments.

A k-partitioning technique takes a collection of n items, D =x_1,x_2 xn, and a positive number K, and divides it into precisely K distinct subsets, D1,∈ D_k,D_k. Set apart with this break. Clustering theory states that objects with similar properties are more closely connected than others. The difficulty of deciding may be reduced by developing a cost function that evaluates the success of clustering for each subset of the dataset. The characteristic of each gene is shown here as an integer. So, the number of characteristics an object possesses may be represented as a row vector of real numbers of length d. For the sake of argument, let's assume that all of the data in the crop yield dataset is complete and that each item has the same qualities. Let there be n objects in the set. $x_i \in D_k$For the sake of brevity, we shall abbreviate the j[th] property of xi as x_{ij}. A D attribute matrix for an object set is denoted by the notation X = (x_{ij}).

$j_G(\Delta) = \sum_{k=1}^{k} \sum_{x_i \in D_k} (x_i - m_k) G (x_i - m_k)$ -------- (1)

$m_k = \frac{1}{n_k} \sum_{x_i \in D_k} x_i$ ------- (2)

G is a positive symmetric weighted matrix, where n_k and m_k are the means and the size of D_k, respectively. A symmetric positive matrix G* meeting Equation (4) is sought via the weighted k-means approach such that the desired subset is indicated by *.

$j_g(\Delta^*) = \min_{\Delta} \{ j_g(\Delta) \}$ ------- (3)

When $j_g(\Delta^*)$ It is computed by multiplying a partition by a weighted matrix G; the output might vary. Thus, it is necessary to normalize the weighted matrix. The G determinant is assumed to be 1 in this investigation.

$\big(det(G)\big) = 1$ ------- (4)

Equation (4) is met because G = I in (4), and the cost function and optimum goal of a typical k-means algorithm are defined by equations (5) and (6), respectively.

Let's say that a collection of data, denoted by$X = x_1, \dots x_n$, exists in a d-dimensional Euclidean space Rd. The k-means approach seeks to minimize an objective function to partition a data set X into a desired number of clusters, k:

$P(U, Z) = \sum_{i=1}^{n} \sum_{l=1}^{k} U_{il} \sum_{j=l}^{d} d(X_{ij}, Z_{lj})$------ (5)

to which $U_{il} \sum_{j=l}^{d} d(X_{ij}, Z_{lj})$, which signifies that the i[th] data point Xi is part of theU_{il}[Th] cluster, and [U] is a [n * k] partition matrix, are binary variables. If Z is a collection of k-vectors representing the cluster centers, then the distance between the i[th] data point and the l[th] cluster centre on the j[th] variable is denoted by d (X_{ij}, Z_{lj}).

3.2.3 Average filling

Average filling refers to a critical crop growth stage such as grains. Seeds gather the bulk of their dry weight, substantially influencing eventual output and quality O. S. Bişkin et al. (2021). This stage includes grain production and growth, nutrition transfer from vegetative components, continued photosynthesis, a sufficient water supply, and appropriate weather circumstances. Monitoring during this stage is critical for yield prediction. It involves assessing crop health, potential, and pest control and employing data and agronomic experience to optimize strategies for increased crop output.

$$GPS(\%) = \frac{TGW_{dry}}{TSW_{dry}} \times 100\% \text{ ------ (6)}$$

The grain percentage of spike weight (on a dry weight basis) is denoted as GPS (%), where TGW_{dry} and TSW_{dry} They are measured in grams, respectively.

$$SMC(\%) = \frac{TSW_{fresh} - TSW_{dry}}{TSW_{fresh}} \times 100\% \text{ ------- (7)}$$

Grain weight (GW) with time for the evaluated winter wheat under varying water and fertilizer availability was the best fit by a sigmoid growth function:

$$GW = GW_{max}\left(1 + \frac{t_e - t}{t_e - t_m}\right)\left(\frac{t}{t_e}\right)\frac{t_e}{t_e - t_m} if 0 \le t \le t_e \text{ ----- (8)}$$

Grain weight (mg), days since a thesis (ds), and plants produced (n) are entered into the equation. Grain weight reaches its maximum value, GW_{max}, at the end of growth, the maximum filling rate emerges at tm.

3.2.4 Decision tree regression

The benefits of employing decision trees are highlighted. Firstly, DTs are easily understandable and interpretable since trees can be visually represented T. Manvitha and K. S. Rekha (2023). Unlike many other methods, DTs need very little preprocessing, such as normalization and standardization of data. Crop yield prediction with DT also uses fewer computer resources than other techniques because of its logarithmic complexity. Predicting both longitude and latitude requires an algorithm able to handle multi-output issues, which is the strength of the DT approach. Data types such as categorical, continuous, ordered, and unordered are all supported by DT, which is a huge plus.

Soft classification outputs for a pixel are commonly scaled from 0 to 1 to more accurately represent the class proportions inside a pixel region on the ground. Therefore, DT(i), where i = 1,..., M, stands for the projected class proportions by the tree i and the normalization of these proportions is as,

$$p(i) = \frac{DT(i)}{\sum_i DT(i)},\ i = 1, \dots, M \text{ ------ (9)}$$

The accuracy of the sorting is assessed once it has been done. Traditional error matrix-based measures are normally reserved for examining the correctness of a hard classification. In contrast, a fuzzy error matrix-based measure may be used to evaluate the efficacy of a soft classification. We evaluate DTR-based soft categorization in this work utilizing the latter two metrics.

Put Q in place of the data at the m[th] node. Separate the information into q_{left} and q_{right} categories for each split = (j, tm) combination of a feature j and a threshold tm.

$$q_{left}(\theta) = (x, y)|x_j <= t_m \text{ ----- (10)}$$

$$q_{right}(\theta) = q \backslash q_{left}(\theta) \text{ ------ (11)}$$

Different impurity functions $H\left(q_{left}(\theta)\right)$ they are used to compute the impurity at m for different problem types (Regression vs classification).

$$G(q, \theta) = \frac{n_{left}}{n_m} H\left(q_{left}(\theta)\right) + \frac{n_{right}}{n_m} H\left(q_{right}(\theta)\right) \text{ ---- (12)}$$

Select the parameters that minimize the impurity

Algorithm 1: Ensemble based Dataset Normalization

Input:

Crop yield dataset, Grain weight at maximum growth (GW_{max}),

Algorithm Steps:

Weighted k-means

Determine the cluster centres$Z = Z_1, Z_2, \dots, Z_n$arbitrarily or according to another scheme.

Initialize cluster centers arbitrarily.

Average filling

Update the partition matrix U based on the assigned data points, considering weights and distances.

For each time point t from 0 to t_e:

a. Calculate grain weight (GW) using a sigmoid growth function.

b. Append calculated GW to the list.

c. If t equalst_m, calculate the maximum grain filling rate (GW_{max}).

d. Calculate the average grain filling rate (AGFR) for the time frame.

Decision tree regression

Train model with the crop yield dataset, using decision trees

Generate an output consisting of the chosen subset of characteristics

Output:

The list of grain weights (GW), the average grain filling rate (AGFR), and the maximum grain filling rate (GW_{max}).

Crop yield data gathering The Ensemble techniques are utilized to do the normalization procedure.

3.3 Ensemble feature selection

After crop yield data normalization, ensemble feature selection techniques isolate the most important features. To achieve a more reliable and complete feature subset, combining different feature selection approaches is necessary. Each feature's significance is calculated using a combination of Logistic Regression (LR) and Random Forest (RF). RF measures the impact of features on model accuracy, while LR provides insights into the individual effects of features on the target variable. Combining the results from both algorithms gives a more reliable ranking of feature importance.

3.3.1 RF with LR

When used in conjunction with approaches such as Random Forest, ensemble feature selection may significantly improve the accuracy and interpretability of crop production prediction models. Using a mix of different decision trees, this technique identifies and prioritizes the most significant characteristics among many elements impacting crop yields, such as weather conditions, soil attributes, and agricultural practices.

Feature selection is another use of Random Forest, whereby a subset of features is chosen from the whole set of crop data input variables S. R. Sani (2023). Reducing dimensionality, increasing model performance, and promoting interpretability are all crucial reasons for feature selection. Each feature's value or relevance is evaluated with the prediction task as part of Random Forest's feature selection process. To determine a feature's relative value, the algorithm considers its effect on the random forest model's predictive power.

The convergence theorem, generalized error, and unconventional estimations form the foundation of a Random Forest. Following is the formula for a random forest:

$$\{h(X, \emptyset_k), k = 1,2, \dots, K\} \text{ --------- (13)}$$

The X-dimensional collection of sample-condition attributes, the k-dimensional baseline classifier parameter, and the s-dimensional sample size:

$$T = \{(x_i, y_i), x_i \in X, y_i \in Y, i = 1,2, \dots, N\} \text{ --------- (14)}$$

X denotes a collection of M-dimensional attribute vectors, whereas Y is the determining factor.

Random forest generalization mistakes are as follows:

$$PE^{*^{def}} = P_{x,y}(av_k I(h(X, \emptyset_k) = Y) \text{-------- (15)}$$

$$-\max_{j=y} av_k I(h(X, \emptyset_k) = j) < 0) \text{--------- (16)}$$

It quantifies how wrong a random forest is in classifying a specific dataset. The following convergence theorem existed during the period K:

$$PE^* \xrightarrow{a.s,} P_{X,Y}(P_o(I(h(X, \emptyset) = j) < 0) \text{ -------- (17)}$$

$$-\max_{j=y} P_o(h(X, \emptyset) = j) < 0) \text{-------- (18)}$$

The generalized error limits of random forests are obtained by combining inequality with equation (28):

$$PE^* \leq \frac{p(1-s^2)}{s^2} \text{-------- (19)}$$

Where s is the basic classifier's accuracy and p is the correlation between the two.

The probabilities of occurrences that may be classified into two groups are predicted using the statistical method of Logistic Regression P. Mishra and R. K. Somkunwar (2023). This supervised learning method is used in many fields, including machine learning, statistics, and medical research. While linear Regression is used to predict continuous values, logistic Regression estimates the probability of a binary outcome based on a collection of continuous predictor factors and a binary target variable. Using the sigmoid function, any real number may be converted into a probability value between zero and one.

A maximum possibility estimate is used to fit a logistic function to the training data, which is how the logistic regression algorithm gets its desired results. The approach optimizes the logistic function's parameters (coefficients) during training such that the discrepancy between the probabilities predicted by the function and the actual binary labels in the training data is as small as possible. Optimization methods like gradient descent are often used for this purpose.

Simple (two-variable) Regression and multiple Regression are both subsets of the basic single-equation linear regression model, which may be written mathematically as

$$Y = a + \sum_{i=1}^{k} b_i x_i + u \text{ ------- (20)}$$

where Y is the result, $Y = x_1, x_2, x_i, \dots, x_k$Are the k independent variables, a and bi are regression coefficients standing in for the model parameters for a given population, and u is a stochastic disturbance term standing in for the effect of unspecified independent variables and a random element in the specified relationship.

3.3.2 Elastic Net with Principal Component Analysis

Specifically, we are talking about the R package glmnet, which implements a technique for fitting generalized linear models (GLMs) using an elastic net penalty A. Lakshmanarao et al. (2023). Each gene's expression value is normalized to zero before being sent to glmnet. After that, we turn off Glmnet's standardized feature. The intercept option of glmnet may be set to true if the distribution of classes in the training set is similar to the predicted distribution in the testing set (that is, they are an accurate prior). If other variables are required, they might be coupled with the genes. The use of dummy variables to denote categories is possible. Rescaling may be necessary to guarantee the mean and standard deviation of continuous variables agrees with gene expression data. Each data point is assigned equal importance by default during the classifier's training. Ensuring the mean and standard deviation alignment with gene expression data might require rescaling for continuous variables. Notably, in the classifier's default training, each data point is assigned equal importance, but it's worth considering weighting to account for variables such as temperature, humidity, pH value, and rainfall.

For the elastic net, 'loss + penalty' is the objective function:

$$\min_{\beta 0,\beta} \frac{1}{N} \sum_{i=1}^{N} w_i l(y_i, \beta_o + \beta^T x_i) + \gamma \left(\frac{(1-\alpha)||\beta||2}{2} + \alpha||\beta||1 \right) \text{ ---- (21)}$$

The symbol w_i stands for the observational value. I l(y) represents the influence on the negative Logarithm of the probability of making an observation. The regularization parameter l (whose functional form is model-specific) calculates the shrinkage, 2 is the L2-norm, 1 is t, and the elastic net penalty sets the weights for ridge and lasso regression. Because no one part is more crucial than any other,

$$\sum_{i=1}^{N} w_i = N \text{ ------ (22)}$$

$$w_i = \frac{M}{n_i}, \text{ ------- (23)}$$

Where the sum of the records in the batch of which i is a member is denoted by n_i.

Since the standard PCA method incorporates all training crop yield dataset in the eigenspace calculation, it does not account for class differentiation F. Peng et al. (2023). Finding the eigenvector might be a challenging intermediary step if there are many training images or the picture

dimensions are high. This is because updating a conventional PCA model with more training dataset requires recalculating the eigenspace, eigenvalues, and feature vectors for each dataset, which is a very inefficient use of computational resources. Adopting a novel training and projection technique has considerably simplified the training process in Superior PCA. To build an eigensubspace and a set of feature parameters, Superior PCA first filters through the training dataset and categorizes the crop dataset. Choose the subject whose eigensubspace best approximates the test dataset.

1. Let the training set of all dataset X can be described as

$X = \{X_1, X_2, X_3 \dots X_L\}$ ------- (24)

2. Compute the mean vector of all training dataset of i[th] person

$X_I = \frac{1}{N_I}\sum_{k=1}^{Ni} X_k^i \ (i = 1,2, \dots l)$ ------- (25)

3. Compute the covariance of the training set of the i[th] person

$S_{x_i} = \frac{1}{N_i}\sum_{k=1}^{Ni}(X_k^i - X_i)$ ------ (26)

4. Compute Matrix *Xi S* m largest eigenvalues *I j u*, where *j* = 1,2...,*m*

Algorithm 2:Ensemble feature selection

Input:

- A training dataset: $X = \{X_1, X_2, X_3 \dots X_L\}$ (dataset)
- Number of dataset for each person: $\{N_1, N_2, N_3 \dots N_L\}(N_i)$
- Number of eigenvalues/features to compute: m

Step:

RF with LR

1. Initialize empty lists: eigenvectors (U), mean vectors (X_i), covariance matrices $(S_{(x\_i)})$
2. Train a random forest with crop yield dataset,

   Elastic Net with Principal Component Analysis
3. Iterate over each person's training set: for I in range(1, l):

   a. Compute the mean vector of the i[th] training dataset: X_i = (1 / N_i) * sum(X_k^i) for k in range(1, N_i)

   b. Compute the covariance matrix of the i[th] crop yield training Dataset: $(S_{(x\_i)})$ = (1 / N_i) * sum$X_I = \frac{1}{N_I}\sum_{k=1}^{Ni} X_k^i \ (i = 1,2, \dots l)$ for k in range(1, N_i)

   c. Compute the eigenvectors (u_i) corresponding to the m largest eigenvalues of $(S_{(x\_i)})$: eigenvalues, eigenvectors = compute eigenvectors $(S_{(x\_i)})u_i$ = eigenvectors[:, :m]

   d. Append u_i, X_i, and $(S_{(x\_i)})$ to the respective lists

   Retrieve the learned coefficients from the fitted model:

   β_0 <- coef(model)["(Intercept)"]

   β <- coef(model)[-1]

   Initialize the coefficients (a and b_i) with random values or zeros
4. Return the lists of eigenvectors (U), mean vectors (X_i), and covariance matrices $(S_{(x\_i)})$

Output:

- Eigenvectors (U) corresponding to the m largest eigenvalues for each person's training set
- Mean vector (X_i) for each person's training set

Covariance matrix $(S_{(x\_i)})$ for each person's training set

Selected features subset

3.4 Classification using SVM

Creating a model to classify crop yields based on various parameters including weather, soil characteristics, and agricultural practices requires

the utilization of Support Vector Machine (SVM) classification. SVM is a robust technique that identifies a hyperplane in the feature space, optimizing the margin between different yield classes. It can also handle non-linear relationships through the use of kernel functions. Prior to implementing SVM, the data undergoes preprocessing and feature manipulation. Preceding SVM implementation, data undergoes preparatory steps encompassing preprocessing and manipulation of features, which encompass variables such as temperature, humidity, pH value, and rainfall.

Incorporating key variables such as relative humidity, pH value of the soil, and rainfall in millimeters (mm) is essential for accurate crop yield prediction. These variables contribute significantly to the yield outcome and should be included in the feature set used for training the SVM model. Relative humidity influences plant growth and water availability, soil pH affects nutrient availability, and rainfall plays a pivotal role in supplying water to the crops. The SVM algorithm can effectively handle these variables as part of its feature space, enabling it to create informed predictions about crop yields based on their relationships with these factors.

The SVM is a nonlinear classifier in the parameter space because the mapping from the input pattern space to the high dimensional feature space is nonlinear. The optimization problem presented by SVM training is quadratic. Here, w is the vector coefficients, and b is the bias factor. It turns out that with huge separation margins, the only thing that matters is how close the points are to each other. The kernel function is used to calculate this kind of similarity. There is no universally accepted procedure for selecting an appropriate kernel function for a given situation.

$D = \{(x^1, y^1), \dots (x^1, y^1)\}, xI\ R^n, y\ I\ \{1, -1\}$ ------- (27)

If the distance between the vectors nearest to the hyperplane is the largest, then the separation produced by the hyperplane is best. A canonical hyperplane is a hyperplane with parameters w and b such that only if these constraints hold,

$\min_i |< w, x^i > +b| = 1$ ------- (28)

Training errors may be kept to a minimum while profit maximization is still possible by adjusting the regularization value 'C'. This is referred to as a "soft margin." Therefore, a kernel function and a regularization parameter are needed to develop a support vector machine.

Sometimes, the SVM will not use a linear boundary but instead will project the input vector x onto a high dimensional feature space z.

An inner product in feature space has an equivalent kernel in input space,

$K(x, x) =< j(x), j(x) >$ -------- (29)

Nonlinear modelling is where a polynomial mapping comes in,

$K(x, x) =< x, x >^d$ ------ (30)

where d is the polynomial degree. There has been a lot of focus on radial basis functions, often using a Gaussian of the type,

$K(x, x) = exp - \frac{||x-x||^2}{2\sigma^2}$ ------- (31)

Algorithm 3: Classification Algorithm

Input:

Crop yield dataset

Desired number of selected features, N

Algorithm:

Train a model with the crop yield dataset

Calculate the predicted probabilities for each instance in the training dataset

$$K(x, x) =< j(x), j(x) >$$

Calculate the error or difference between the predicted probabilities and the actual

$$K(x, x) =< x, x >^d$$

Output:

The SVM classification model predicts Crop yield

Finally, the classification was completed successfully with the help of a Support Vector Machine (SVM), a popular machine learning technique recognized for its resilience and ability to handle difficult decision boundaries. SVMs are highly useful in fields such as image classification, text classification, and bioinformatics, and they are well-known for their high performance on small to medium large datasets.

IV. Results And Discussion

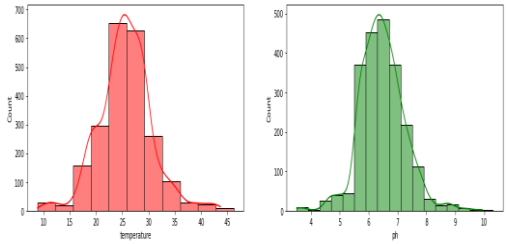

Figure 2: Temperature analysis

The temperature analysis may show in Figure 2. The x-axis represents the temperature, while the y-axis displays the count. X-axis: The x-axis represents the year's temperature. The y-axis represents the count or quantity of something being measured or observed.

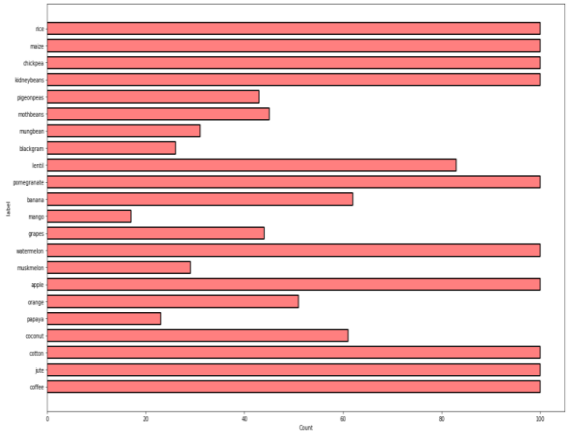

Figure 3: Plot graph

Figure 3 depicts a graph where the X-axis represents counts of a certain variable, and the Y-axis represents corresponding labels. The X-axis values could represent discrete quantities, time intervals, or any form of numerical count, while the Y-axis labels provide context or meaning to these counts. Each point on the graph then represents a count-label pair, visually illustrating the relationship between the numerical counts and their associated labels. The graph may have additional elements like a title, axis labels, and data points to enhance its clarity and interpretability.

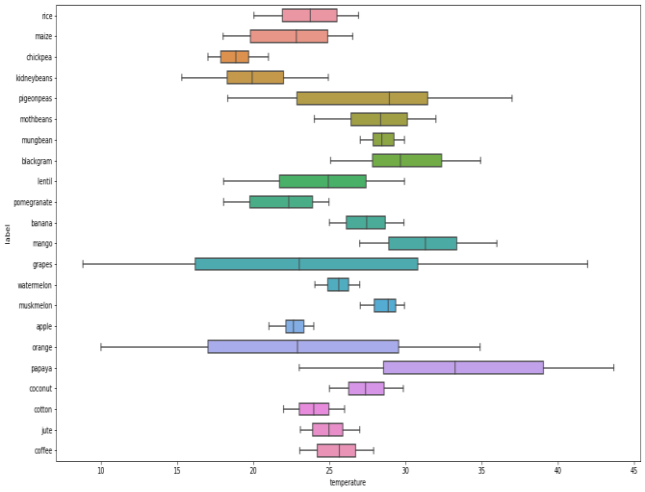

Figure 4: Feature selection

In Figure 4, which illustrates feature selection, the X-axis represents temperature values, while the Y-axis displays corresponding labels. Specifically, the X-axis is likely to represent a range of temperature values, which could be discrete data points or a continuous scale.

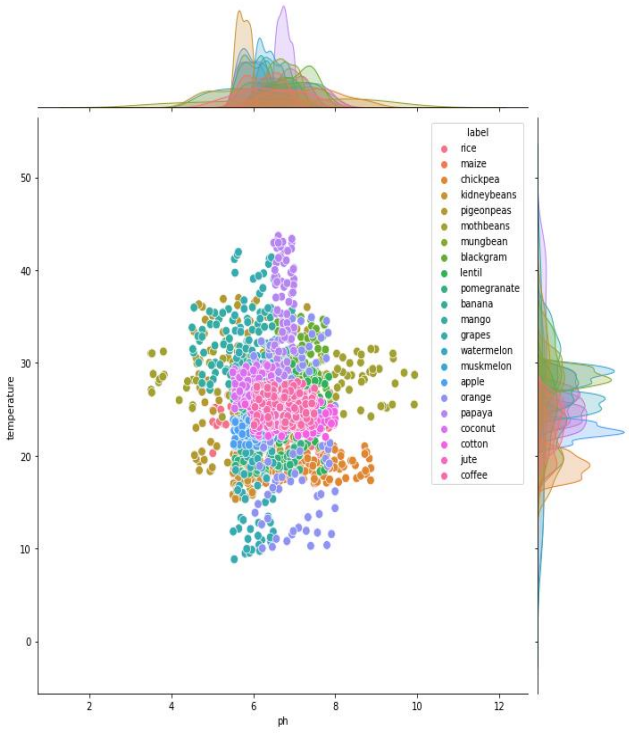

Figure 5: Feature embeddings

Figure 5 shows feature embedding. The x-axis shows ph, and the y-axis shows temperature.

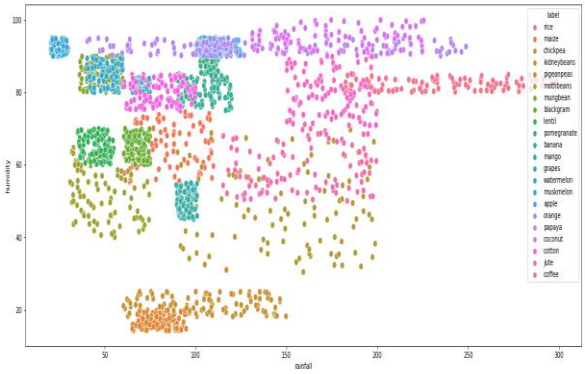

Figure 6: Climate Analysis

In Figure 6, representing a climate analysis, the X-axis represents rainfall values, while the Y-axis portrays humidity levels. The X-axis likely encompasses a range of rainfall measurements, indicating different precipitation amounts.

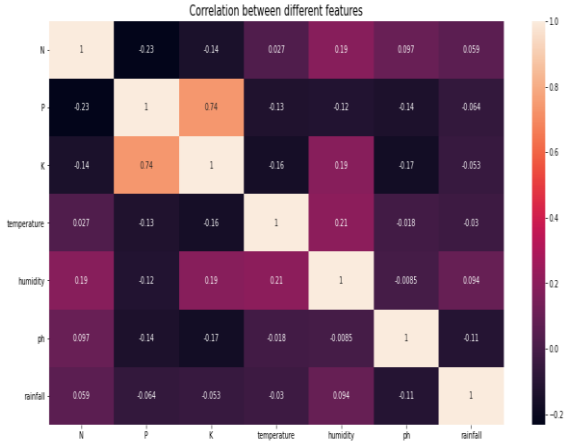

Figure 7: correlation metrics

Figure 7 shows correlation metrics. Correlation metrics are statistical tools used to quantify the strength and nature of relationships between variables. The Pearson correlation coefficient gauges linear associations, ranging from -1 (perfect negative correlation) to 1 (perfect positive correlation), while Spearman's rank correlation and Kendall's Tau evaluate monotonic relationships.

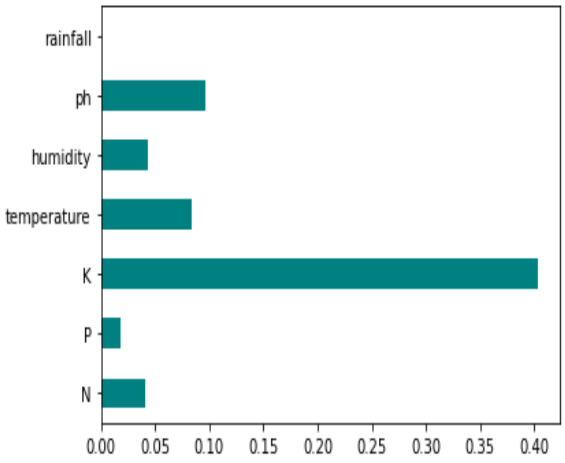

Figure 8: Important Feature selection

As can be seen in Figure 8, the best features have been chosen using the ensemble feature selection. The features and reviews as a whole. There has had a significant effect on the title. A global accuracy of 0.91 per cent has been reached.

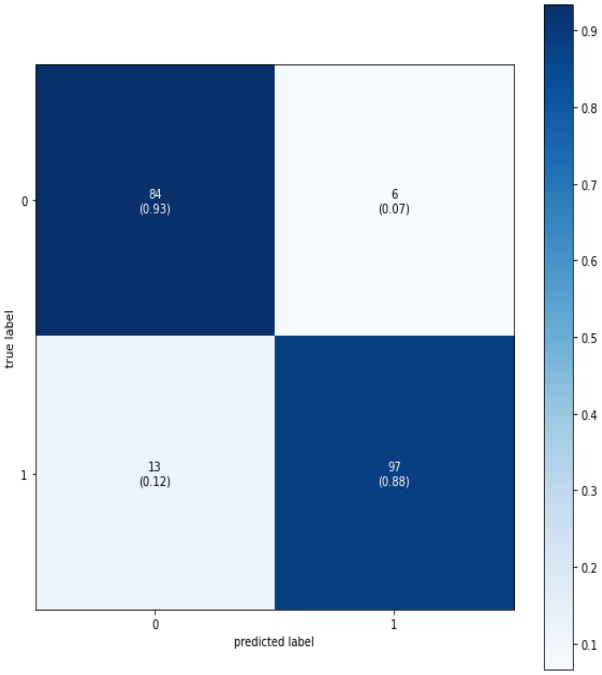

Figure 9: SVM confusion metrics

Figure 9 shows SVM confusion metrics TP, FP, TN, and FN values are represented in Figure 10. The predicted class for TP is 84, TN is 97, and FP and FN are 6 and 13

Table 1: performance metrics comparison

$$\text{Accuracy} = \frac{\text{True Positive+True Negatives}}{\text{False Positives+False Negatives+True Positive+True Negatives}} \text{ -------- (41)}$$

$$\text{Precision} = \frac{\text{True positive}}{\text{True positive+False positive}} \text{ ---------- (42)}$$

$$\text{Recall} = \frac{\text{True positive}}{\text{True positive+False Negative}} \text{ -------- (43)}$$

$$\text{Fmeasure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision+Recall}} \text{ ------- (44)}$$

	CNN	Decision tree	SVM
Accuracy	0.84	0.88	0.91
Precision	0.86	0.90	0.94
Recall	0.81	0.85	0.88
F-measure	0.83	0.87	0.91

Table 1 shows the three machine learning models evaluated; Support Vector Machine (SVM) exhibits the strongest performance across all metrics. With an accuracy of 0.91, SVM outperforms Convolutional Neural Network (CNN) and Decision Tree models, showcasing its ability to classify instances accurately. Additionally, SVM's higher precision (0.94) signifies its proficiency in correctly predicting positive instances, while its recall (0.88) demonstrates its capability to capture actual positive instances effectively. The balanced F-measure of 0.91 further reinforces SVM's well-rounded performance, making it a favourable choice based on this evaluation.

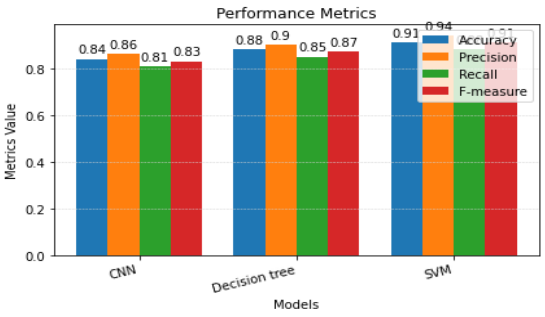

Figure 10: performance metrics comparison

Figure 10 shows a performance metrics comparison chart. The x-axis shows models, and the y-axis shows metrics values.

V. Conclusion

In conclusion, this paper presents a comprehensive approach to crop yield prediction, highlighting the significance of accurate data integration, preprocessing, feature selection, and classification methods. The study emphasizes the importance of data normalization and feature selection in improving agricultural production projections. By employing these techniques, we aim to improve the performance and interpretability of machine learning models. The weighted K-means clustering and average filling techniques ensure a more accurate dataset representation by appropriately considering sample weights and handling missing values. This normalization step prepares the dataset for subsequent analysis. Combining Random Forest, Logistic Regression, and Elasticnet, the ensemble feature selection approach allows us to identify the most relevant features by considering their importance, relevance, and coefficients. We have achieved an accuracy of 0.91%. The integration of diverse data sources, encompassing attributes such as rainfall, humidity, and nutrient levels (N, P, K), reflects the complexity of factors influencing crop health and growth. This helps in reducing dimensionality and eliminating noise or redundant information. Principal Component Analysis further enhances the feature selection process by reducing the dimensionality of the dataset while preserving its important characteristics.

Reference

[1] A. Lakshmanarao, M. N. Kumar, K. S. V. Ratnakar and Y. Satwika, "Crop Yield Prediction using Regression Models in Machine Learning," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 423-426, doi: 10.1109/ICAAIC56838.2023.10141462.

[2] Araque, O., Corcuera-Platas, I., Sánchez-Rada, J. F., & Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. Expert

Systems with Applications, 77, 236–246. doi:10.1016/j.eswa.2017.02.002

[3] Banerjee, R., Marathi, B., & Singh, M. (2020). Efficient genomic selection using ensemble learning and ensemble feature reduction. Journal of Crop Science and Biotechnology. doi:10.1007/s12892-020-00039-4

[4] Conţiu, Ş., & Groza, A. (2016). Improving remote sensing crop classification by argumentation-based conflict resolution in ensemble learning. Expert Systems with Applications, 64, 269–286. doi:10.1016/j.eswa.2016.07.037

[5] Elghazel, H., & Aussem, A. (2013). Unsupervised feature selection with ensemble learning. Machine Learning, 98(1-2), 157–180. doi:10.1007/s10994-013-5337-8

[6] F. Peng, M. Guo, C. Zheng, S. Wang, X. Wang and M. Xu, "An Assessment Model of Digital Literacy for the Students in Vocational Education Based on Principal Component Analysis in Machine Learning," 2023 IEEE 6th Information Technology, Networking, Electronic and Automation Control Conference (ITNEC), Chongqing, China, 2023, pp. 1382-1386, doi: 10.1109/ITNEC56291.2023.10082530.

[7] Fayyazifar, N., &Samadiani, N. (2017). Parkinson's disease detection using ensemble techniques and genetic algorithm. 2017 Artificial Intelligence and Signal Processing Conference (AISP). doi:10.1109/aisp.2017.8324074

[8] Gaikwad, D. P., & Thool, R. C. (2015). Intrusion Detection System Using Bagging Ensemble Method of Machine Learning. 2015 International Conference on Computing Communication Control and Automation. doi:10.1109/iccubea.2015.61

[9] J. Dan, "Research and Improvement of K-means Clustering Analysis Algorithm in the Information Warfare," 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2022, pp. 284-287, doi: 10.1109/ICCSMT58129.2022.00066.

[10] Kaur, I., & Kaur, A. (2021). A Novel Four-Way Approach Designed With Ensemble Feature Selection for Code Smell Detection. IEEE Access, 9, 8695–8707. doi:10.1109/access.2021.3049823

[11] Laradji, I. H., Alshayeb, M., &Ghouti, L. (2015). Software defect prediction using ensemble learning on selected features. Information and Software Technology, 58, 388–402. doi:10.1016/j.infsof.2014.07.005

[12] Moghimi, A., Yang, C., & Marchetto, P. M. (2018). Ensemble Feature Selection for Plant Phenotyping: A Journey from Hyperspectral to Multispectral Imaging. IEEE Access, 1–1. doi:10.1109/access.2018.2872801

[13] N. Omar, A. Al-zebari and A. Sengur, "Improving the Clustering Performance of the K-Means Algorithm for Nonlinear Clusters," 2022 4th International Conference on Advanced Science and Engineering (ICOASE), Zakho, Iraq, 2022, pp. 184-187, doi: 10.1109/ICOASE56293.2022.10075614.

[14] O. S. Bişkin, T. Saydam and S. Aksoy, "The Averaging Effect on Resonant Frequency Calculations of a Partially Filled Microwave Cavity Using FDTD Method," 2021 IEEE Microwave Theory and Techniques in Wireless Communications (MTTW), Riga, Latvia, 2021, pp. 129-133, doi: 10.1109/MTTW53539.2021.9606857.

[15] P. Mishra and R. K. Somkunwar, "Smart Irrigation with Water Level Indicators Using Logistic Regression," 2023 4th International Conference for Emerging Technology (INCET), Belgaum, India, 2023, pp. 1-5, doi: 10.1109/INCET57972.2023.10170388.

[16] Pham, B. T., Nguyen-Thoi, T., Qi, C., Phong, T. V., Dou, J., Ho, L. S., … Prakash, I. (2020). Coupling RBF neural network with ensemble learning techniques for landslide susceptibility mapping. CATENA, 195, 104805. doi:10.1016/j.catena.2020.104805

[17] Prusa, J. D., Khoshgoftaar, T. M., & Napolitano, A. (2015). Using Feature Selection in Combination with Ensemble Learning Techniques to Improve Tweet Sentiment Classification Performance. 2015 IEEE 27th International Conference on Tools with Artificial Intelligence (ICTAI). doi:10.1109/ictai.2015.39

[18] Rai, A. (2020). Optimizing a New Intrusion Detection System Using Ensemble Methods and Deep Neural Network. 2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184). doi:10.1109/icoei48184.2020.9143028

[19] S. R. Sani, S. V. Sekhar Ummadi, S. Thota, N. Muthineni, V. S. Srinivas Swargam and T. S. Ravella, "Crop Recommendation System using Random Forest Algorithm in Machine Learning," 2023 2nd International Conference on Applied Artificial Intelligence and Computing (ICAAIC), Salem, India, 2023, pp. 501-505, doi: 10.1109/ICAAIC56838.2023.10141384.

[20] Safiyari, A., & Javidan, R. (2017). Predicting lung cancer survivability using ensemble learning methods. 2017 Intelligent Systems Conference (IntelliSys). doi:10.1109/intellisys.2017.8324368

[21] Shorewala, V. (2021). Early detection of coronary heart disease using ensemble techniques. Informatics in Medicine Unlocked, 100655. doi:10.1016/j.imu.2021.100655

[22] T. Manvitha and K. S. Rekha, "Improved Accuracy for prediction of leaf wetness using Logistic Regression algorithm compared with Decision Tree algorithm," 2023 Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2023, pp. 1-5, doi: 10.1109/ICONSTEM56934.2023.10142550.

[23] Tajik, S., Ayoubi, S., &Zeraatpisheh, M. (2020). Digital mapping of soil organic carbon using ensemble learning model in Mollisols of Hyrcanian forests, northern Iran. Geoderma Regional, e00256. doi:10.1016/j.geodrs.2020.e00256

[24] Tan, M., Yuan, S., Li, S., Su, Y., Li, H., & He, F. (2020). Ultra-short-term industrial power demand forecasting using LSTM based hybrid ensemble learning. IEEE Transactions on Power Systems, 1–1. doi:10.1109/tpwrs.2019.2963109

[25] Tang, Z., Wang, H., Li, X., Li, X., Cai, W., & Han, C. (2020). An Object-Based Approach for Mapping Crop Coverage Using Multiscale Weighted and Machine Learning Methods. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 1–1. doi:10.1109/jstars.2020.2983439

[26] Verma, A. K., Pal, S., & Kumar, S. (2019). Comparison of skin disease prediction by feature selection using ensemble data mining techniques. Informatics in Medicine Unlocked, 100202. doi:10.1016/j.imu.2019.100202

[27] Weng, B., Lu, L., Wang, X., Megahed, F. M., & Martinez, W. (2018). Predicting short-term stock prices using ensemble methods and online data sources. Expert Systems with Applications, 112, 258–273. doi:10.1016/j.eswa.2018.06.016

[28] Yekkala, I., Dixit, S., & Jabbar, M. A. (2017). Prediction of heart disease using ensemble learning and Particle Swarm Optimization. 2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon). doi:10.1109/smarttechcon.2017.8358460

[29] Zheng, Y., Li, G., Zhang, W., Li, Y., & Wei, B. (2019). Feature Selection with Ensemble Learning Based on Improved Dempster-Shafer Evidence Fusion. IEEE Access, 1–1. doi:10.1109/access.2018.2890549

[30] Zhou, Z., Li, Y., Zhang, Q., Shi, X., Wu, Z., & Qiao, Y. (2015). Comparison of Ensemble Strategies in Online NIR for Monitoring the Extraction Process of Pericarpium Citri Reticulatae Based on Different Variable Selections. Planta Medica, 82(01/02), 154–162. doi:10.1055/s-0035-1558085