A Hybrid Feature Extraction and Feature Selection Mechanism to Predict Disease in Plant Leaves

Abisha A.* and Bharathi N.

Department of Computer Science and Engineering, SRM Institute of Science and Technology, College of Engineering and Technology, Vadapalani Campus, Chennai, India
Email: aa7111@srmist.edu.in (A.A.); bharathn2@srmist.edu.in (B.N.)
*Corresponding author

***Abstract*—The health of the plants is vital to meet the demands of the food cycle. As the symptoms of disease or infection are most commonly seen in plant leaves, selecting features from plant leaves that are highly impacting plant health is crucial. Plant health is a global imperative for food security and ecological balance and must be treated as the top priority. Feature Extraction (FE) and Feature Selection (FS) are significant in Deep Learning (DL) and Machine Learning (ML) models, which are used for classification and prediction. Xception-based feature extraction and random forest classification yield accurate predictions, offering interpretability and adaptability across diverse plant diseases and datasets, benefiting agriculture. In this article, FE is performed using an Xception pre-trained model and the extracted features are sent for FS. Further, six FS methods such as ANOVA, chi-square, Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Lasso and Ridge, have been deployed and compared with machine learning algorithms such as Logistic Regression (LR), K Nearest Neighbours (KNN), Decision-Trees (DT), Random Forest (RF), Support Vector Machine (SVM), Naive-Bayes (NB) for classification. The article also proposes an Ensemble Feature Selection (EFS)-RF method, which combines feature sets from six feature selection algorithms and classifies based on majority voting. The methodology section details criteria for selecting FE and FS methods, utilizing an ensemble to maximize their respective benefits. The paper contributes to agriculture by employing a hybrid approach, integrating DL (Xception-based FE) and ML (RF-based Classification), utilizing an ensemble of FS methods to identify and assign higher weightage to features prevalent across subsets. The proposed method has outperformed other algorithms for both datasets with 98 % accuracy and 0.02 Mean Squared Error (MSE) for dataset I and 98.125 % accuracy and 0.01875 MSE for dataset II.**

***Keywords*—feature extraction, feature selection, filter, wrapper, embedded, machine learning, ensemble**

I. Introduction

Plants feed animals and humans and assist in preserving the ecosystem and atmosphere by producing oxygen. Abisha and Bharathi [1] explore biotic and abiotic plant stress. Plants provide oils, fuel, fibres, insecticides, medications, colours, timber, and rubber. Various approaches and algorithms are used to evaluate plant health; therefore, identifying the methodology is vital. Plants are stressed by water, salinity, dust, other environmental conditions, and fungi, bacteria, and viral illnesses that cause rust, blight, rots, canker, etc. It causes inadequate yield, irreparable damage, wilting, and plant death. Viruses, fungi, bacteria, and insects stress plants. Such creatures drain plant nutrients, shortening their lives. Biological stress causes preharvest and postharvest losses. Abiotic stress is also a factor that affects plant health. Chakraborty and Newton [2] climate change affects plant health, resulting in production, quality and food security. Though many articles and research have evolved, improving the performance to more accurate and promising results is essential. Hence, we have introduced an ensemble method.

Feature Selection (FS) techniques should be eminent from Feature Extraction (FE). In contrast to feature selection, which produces a subset of the features, Feature Engineering (FE) creates new features by deriving current features' functions in the system. This research aims to investigate the feature selection approaches compatible with machine learning classification, and a suggested ensemble method will be presented. A wide variety of algorithmic approaches to feature selection have developed throughout time. They may be broken down into one of three categories. Filter-based, wrapper-based, and embedded feature selection algorithms are the three groups that fall under the broad heading of feature selection. Comparisons are made between the Feature Selection (FS) approaches and the machine learning algorithms that are applied to them.

The classification of plant health is carried out to determine if the plant in question is hale, hearty, or diseased. Machine Learning (ML) models [6] are used to train the feature that is selected. The model is trained using a selection of six different machine-learning techniques. In most cases, eighty to ninety per cent of the available data is used for training. Throughout the training of machine learning, supervised learning and unsupervised learning

Manuscript received July 19, 2023; revised September 6, 2023; accepted October 27, 2023; published April 9, 2024.

are typically the two kinds of learning methods that are used. Supervised learning is the more prevalent of the two, although unsupervised learning is becoming more popular. During the training phase of supervised learning, correct answers are provided as input, and then the learner is taught. In this context, every one of the six algorithms relies on supervised learning methods, and every machine learning model is effective compared to conventional classification models that are analogous to pattern-matching algorithms.

Many FS methods could be used with ML to select features for plant disease recognition. Though many studies have been carried out for plant health recognition, there needs to be a proper study that focuses on comparing all the methods, such as filter, embedded and wrapper for plant health recognition. Hence, there needs to be adequate evidence, and it is unidentified which FS method will work with the ML algorithm for plant health identification. Therefore, this paper compares and proposes an ensemble method for plant disease recognition.

II. Literature Review

A review of nature-based algorithms is presented in [3] for FS. Swarm intelligence-based, and evolutionary algorithms are the primary categories used to cast and classify nature-inspired techniques. These classifications were based on their purpose. Aich *et al.* [4] described supervised feature selection strategies that use various FS strategies. Cisotto *et al.* [5] present a new way to identify the most relevant components to identify gripping tasks. This simplifies recording to reduce internet data transfer and gives physiologically significant elements for medical interpretation. Consensus clustering for feature selection and layered cross-validation for testing enhance algorithm durability. FeSC's robust feature selection and classification architecture is limited by dataset size. Future research could determine if FeSC's size influences its performance in this and other applications. In Ref. [6], 14 active VEOs are accountable for a violent act based on 14 criteria, including human and structural tolls, target kind and value, intelligence, and weapons used. Top-ranked attributes linked to target kind and plan and multilayer perceptron reached 40% test accuracy. Jiang *et al.* [7] integrated the filter and wrapper approaches. The filter approach uses the linear correlation coefficient. Wrapper approach classification uses Support Vector Machine (SVM). After examining the relationship between features and arrhythmias, viable heartbeat feature subsets were found and adopted to the most sensitive one. Different feature lists for each heartbeat improve accuracy and reduce computing burden. Rado *et al.* [8] tested three feature selection strategies on seven numerical and mixed healthcare datasets. Classification performance has been evaluated using 10-fold cross-validation. Experiments indicated that feature selection approaches affect classification performance differently. Saw and Myint [9] used Particle Swarm Optimisation (PSO) to increase the accuracy of healthcare data classification. The WFS-PSO technique improves classification accuracy when used with different algorithms.

In Ref. [10], some of the commonly used assessment measures for feature selection are investigated. Additionally, supervised, unsupervised, and semi-supervised feature selection approaches are reviewed. Finally, the methodology is applied to ML problems such as classification and clustering, and the authors discuss the potential challenges for feature selection. In Ref. [11], a comprehensive assessment of several algorithms for feature selection is offered together with computation. This study also categorises the contributions made by all of the algorithms.

In Ref. [12], an exhaustive study on using Artificial Intelligence (AI) methods to forecast per capita health spending in Turkey is presented. To predict per capita Health Care Expenditures (pcHCE), well-known AI approaches have been used. These techniques include RF, Artificial Neural Network (ANN), Support Vector Regression, and Relevance Vector Machine (RVM). To determine the optimal and the subset of features for the estimate of pcHCE, every approach has been run through a feature selection method called Global Alliance for Food Security (GAFS), which is based on genetic algorithms. In contrast, the findings demonstrated that the GAFS approach improved the overall performance of fundamental AI models by an association that received a score of 99.78% R2. Spencer *et al.* [13] used of ML techniques is carried out using linked characteristics chosen from various FS methods. Additionally, the models are implemented. To develop distinctive features, Principal Component Analysis (PCA), chi-square testing, ReliefF, and symmetric ambiguity were utilised to estimate four commonly used data sets on heart conditions. In this study, several classification algorithms were the first cast to develop ways. These techniques were then distinguished to locate the optimum groups of attributes to advance the classification of right heart disease. The method used in the cardiac records we investigated had an accuracy of 85%.

Yu [14] investigated and researched a fusion feature choice scheme referred to as HTTP File Server (HFS). This scheme was established using the grouping of Gaussian mixture models, and K-means is applied. The unsupervised learning method is used to construct the task that is planned. A Self-Organising Map (SOM), has been used and tested experimentally on behaviour test beds by forecasting and projecting the outcomes of the tests. An innovative health evaluation method known as the Log-Likelihood Probability (LLP) is presented as an understandable recommendation to quantify the well-being of system circumstances and developed by the authors. A phishing detection system based on machine learning is described using the Hybrid Set Feature Selection (HEFS) approach [15]. This method employs a unique feature selection method. It does this using two stages, with the first phase presenting a new method and the second phase picking a particular group of characteristics.

Metselaar *et al.* [16] used openly accessible mRNA expression data in ninety-three patients and twenty-five healthy controls to perform Recursive Ensemble Feature

Selection (REFS). They developed a sign of twenty-three genes capable of detecting distinct cases and controls. REFS beat all other techniques with an Area under Curve (AUC) of 0.92, making it the winner. Hashemi *et al.* [17] suggested a Pareto-based Ensemble of Feature Selection (PEFS) approach that applies a modelled bi-objective optimisation strategy to identify non-dominant features based on a decision matrix created by different FS techniques. This method was developed to locate non-dominant features based on a decision matrix. The non-dominated features are then reorganized in the bi-objective space according to the crowding distance in the next step. The algorithm's requirements order the process of looking for characteristics not dominated by other features.

In Ref. [18], a categorisation backward feature selection technique that is based on ranking information (SBFS-RI), as well as an original ensemble FS method that integrates various ranking information (FS-MRI), is proposed. This method may generate an intuitive threshold value while considering the algorithms' reaction, which would result in the production of the most precise and consistent feature subset.

Pardo *et al.* [19] outlined two distinct feature selection ensemble designs. Both of these designs make use of a variety of different individual approaches. The dataset is then spread over many nodes to cut down on the time needed for computation by parallelizing the training effort. The homogeneously distributed ensemble is then constructed using the same feature selection method. The goal of the heterogeneous centralized ensemble is to take advantage of the individual techniques' strengths while overcoming their limitations. This is accomplished using various feature selection methods on the same training data. The latter approach has the additional benefit of freeing the user to decide whether the technique is more suitable for a particular scenario.

Makimoto *et al.* [20] investigated the influence of many different combinations of FS and classification algorithms to see which combination achieves the highest accuracy when categorising Chronic Obstructive Pulmonary Disease (COPD) status. In addition, Makimoto *et al.* [20] determines the influence that data cleaning had and which texture-based radiomic feature set was the most relevant for COPD categorisation. The research has assumed that texture-based radiomic features can extract information about disease heterogeneity that can detect structural changes and that these features, combined with ML models, will have higher accuracy than traditional methods.

The K-means clustering approach is used for image segmentation in [21], while the Gray-Level Co-occurrence Matrix (GLCM) algorithm is used for FE. The Bayesian Data Analysis (BDA) optimisation technique is utilised for feature selection, and the Extreme Learning Machine (ELM) algorithm is finally used for disease classification in plant leaves. The method that is provided here optimises the input weights as well as the hidden biases for ELM. The dataset utilised in this research consists of seventy-three photos of plant leaves. Because of this, testing has been done on four diseases that often afflict plants. According to the findings of the experiments, the proposed method has achieved encouraging results in terms of these classification measures.

Ensemble FS is treated as a Multi-Criteria Decision-Making (MCDM) method for the first time in [22]. To this end, the well-known MCDM algorithm VIKOR ranks the features by assessing several feature selection strategies as varied decision-making criteria. First, with the help of the rankings of each feature according to different rankers, the suggested technique, EFS-MCDM, creates a decision matrix. When the decision matrix is complete, the authors use the VlseKriterijumska Optimizacija I Kompromisno Resenje (VIKOR) method to score each feature. In the end, the user may choose as many characteristics as they want and a rank vector for them will be generated as an output. It is compared with a few ensemble feature selection approaches to demonstrate the superiority and efficacy of the suggested approach. The results indicate that the strategy outperforms competing methods and does it in a fraction of the time.

Using a mixture of ML base learners, Yaghoobi *et al.* [23] propose arm volume predictions for people with arm lymphedema. Using a Genetic Algorithm (GA), the hyperparameters of the Evolutionary Ensemble Feature Selection Learning (EEFSL) model are improved to boost the model efficiency. Different base learner weights and feature details for each base learner are included in the hyperparameters. The fitness function of GA is to maximise the agreement between anticipated and measured arm volumes. The suggested approach successfully measured sixty arms from 30 people with arm lymphedema. The findings validate the proposed a Horizontal-Vertical Image Scanning-Evolutionary Ensemble Feature Selection Learning (HVIS-EEFSL) approach as a valid and reliable alternative to Water Displacement (WD) and Circumferential Measurement (CM) for measuring arm volume in lymphedema patients.

Dataset examples from the data exfiltration and keylogging subcategories of the Information Theft category have been used to train prediction models using Bot-IoT developed by Leevy [24], which can detect assaults. To this end, the role focuses on determining the impact of ensemble FS strategies (FSTs) on classification efficiency concerning these particular assault examples. The best individual technique won't consistently outperform a collection or ensemble of FSTs. The area under the Precision-Recall Curve (AUC) and the area under the receiver Operating Characteristic Curve (ROC) are two procedures deployed to Determine the Efficacy of a Classification System (AUPRC). The suggested ensemble FSTs are helpful in this study, not because they alter classification performance but because feature reduction reduces computing costs and enhances data presentation, leading to better insights.

Thus, this section has explained various research related to feature selection. Most of the study doesn't determine how many k features would be ideal for FS. This paper focuses on finding the best k value, which would have iterations till all the feature set combinations are made and find the best feature set(K). The contribution and novelty of this work lie in its development of a hybrid approach,

Xception-EFS-RF, which integrates feature extraction using the Xception model, feature selection through various methods, and classification with Random Forests for plant disease prediction. This novel combination leverages the strengths of deep learning, feature selection, and ensemble learning techniques to enhance accuracy and interpretability. Furthermore, the systematic evaluation of multiple feature selection methods and machine learning algorithms across different datasets adds valuable insights into their comparative performance for plant health prediction tasks. This work not only provides an effective predictive model but also offers guidance on selecting suitable feature selection techniques in similar applications, thus contributing to agricultural disease monitoring and management.

III. Materials and Methods

The steps involved in applying the Xception pre-trained model and feature selection methods to the plant health prediction are image pre-processing, feature extraction, variance thresholding, ensemble feature selection and classification. They are described in detail in the below sections.

A. Data

The plant village dataset [25] and the banana leaf image dataset have been compiled by gathering data from various open-source platforms and merging them into a single dataset. The banana leaf dataset consists of images combined from the datasets [26, 27]. The image dataset I for the PlantVillage dataset has a total of 2,000 images with 1000 healthy and 1000 diseased images, while the image dataset II for banana leaves has 1,600 images with 800 healthy and 800 diseased images. There is no class imbalance as both classes have an equal number of images in both datasets. The training and testing sets were split manually. This manual splitting process involves organising your image data into separate directories for training and testing. The split ratio is 80:20. Table I describes the dataset with its associated data.

TABLE I. Dataset Description

	Name	Number of images	Number of training images	Number of testing images	Image size
Dataset I	Plant village dataset	2,000	1,600	400	256×256
Dataset II	Banana leaf images	1,600	1,280	320	256×256

The work involves several modules, including image pre-processing, Feature Extraction (FE), Feature Selection (FS), and classification. Each of these modules is described in detail below. The overall workflow diagram of the paper is shown in Fig. 1.

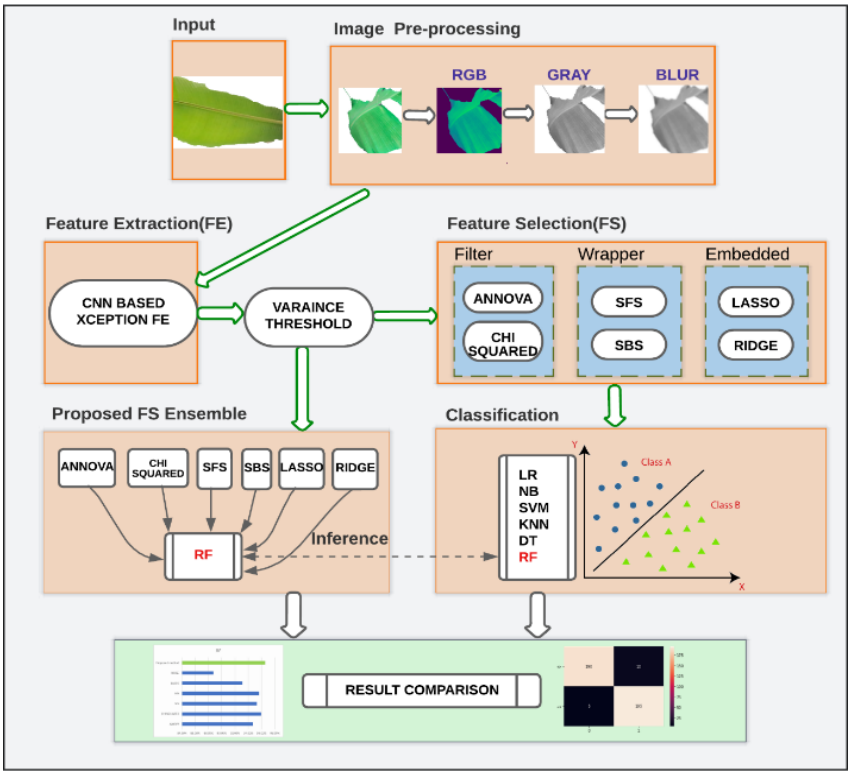

Fig. 1. The overall workflow of the proposed system.

B. Image Pre-processing

The conversion of images is the first stage in pre-processing data. Vishnoi *et al*. [28] discussed many image-processing techniques and the importance of plant disease. Each image is scaled down to a predetermined size of 256×256 pixels. The image processing workflow involves four key steps: firstly, images are initially read in the BGR color model using OpenCV imread function, but they can be converted to the Red, Green, Blue (RGB) color model

if needed through cvtColor. It's crucial to note that Red, Green, Blue (RGB) may introduce blurriness, potentially impacting disease detection and making it unsuitable for pre-processing. Secondly, the RGB color model organises primary colors (Red, Green, Blue (RGB) to create a wide range of colors. Grayscale conversion, the third step, offers advantages in processing efficiency, edge detection, and memory consumption compared to color images. Lastly, image blurring, particularly Gaussian blur, is employed to remove outlier pixels or noise using a low-pass filter. This blurring step is common before implementing techniques like edge detection or contour localisation in various image processing tasks.

C. Feature Extraction

Feature extraction is the process of extracting features from images, as explained by Abisha and Bharathi in [29] and [30] and also in another work by Nandhini and Bhavani [31]. Here, features are extracted using a model called Xception, which is a pre-trained Convolutional Neural Network (CNN) model where features are extracted automatically with various layers like input and hidden layers, as in Abisha and Bharathi [30] features are extracted using this method.

The Xception model, which excludes the dense (fully connected) layers, is renowned for its depth and robust architecture, comprising a total of 132 layers. This deep convolutional neural network has been meticulously designed to excel in the realms of image classification and feature extraction. The layers within this model encompass an array of convolutional layers, each serving a distinct purpose, as well as separable convolution layers that further enhance its efficiency. Additionally, these layers are complemented by an assortment of activation functions, batch normalisation techniques, and other fundamental building blocks commonly associated with convolutional neural networks. It's worth noting that the architecture of deep learning models can exhibit variations, but Xception stands out for its exceptional depth and the remarkable efficiency it demonstrates in capturing intricate features within images.

D. Variance Threshold

The variance threshold serves as a straightforward initial method for feature selection, aiming to eliminate features with insufficient variance by applying a predetermined threshold. Typically, it eliminates features that exhibit zero variance, meaning they possess identical values across all samples as the default behaviour. In this paper, the features obtained after FE had many of the same 0-valued subsets. Hence, a variance threshold was applied to remove the unnecessary features.

E. Feature Selection

Feature selection can be described as, in a set of A features, the part of feature selection is to select a subset of the feature of size B with reduced features (B < A). In this paper, we have used filter-based, wrapper-based and embedded methods for FS. We have implemented Analysis of Variance (ANOVA), Chi-Square, Sequential Forward Selection (SFS), Sequential Backward Selection (SBS), Least Absolute Shrinkage and Selection Operator (LASSO), Ridge feature selection methods and random forest for classification. ANOVA and Chi-Square are effective for selecting relevant features in plant health prediction, with ANOVA suitable for numerical and Chi-Square for categorical data. Sequential Forward Selection (SFS) and Sequential Backward Selection (SBS) help identify informative feature subsets, simplifying models. Lasso and Ridge regularisation techniques are valuable for automatic feature selection and mitigating multicollinearity, respectively, enhancing model performance and stability. The choice of method depends on data type and modelling goals, allowing us to pinpoint critical plant health indicators while improving model interpretability. In the realm of plant disease prediction, the selection of six distinct feature extraction methods is a deliberate and strategic choice. Each method serves a specific purpose, contributing to the robustness of the ensemble approach. ANOVA, for instance, meticulously identifies high-variance features by scrutinising group means, shedding light on features with significant variability. On the other hand, Chi-Square enters the equation to gauge feature relevance, assessing their independence from the target variable with precision. The Sequential Feature Selection methods, including SFS and SBS, add a systematic layer to the process, exhaustively exploring feature subsets to optimise ensemble performance. Meanwhile, LASSO, through its introduction of an L1 penalty, distinguishes itself by effectively selecting relevant features and suppressing irrelevant coefficients. Conversely, Ridge, by employing an L2 penalty, stabilises feature coefficients, creating a well-balanced and robust set. These six meticulously chosen methods amalgamate their outcomes into an ensemble feature set, carefully curated and meticulously prepped. This feature set is then harnessed to train ensemble models like Random Forest and Gradient Boosting, capitalising on the combined wisdom of these methods. The result is an ensemble approach that excels in plant disease prediction, bolstered by rigorous evaluation and cross-validation, promising the utmost in accuracy and reliability.

1) Filter methods

Filter methods can be used for feature selection as well as to pre-process the data in the initial phase of Wrapper Methods (WFS). It is used to remove empty values, remove data that are more correlated to one another, remove redundant data, etc.

a) Analysis of Variance (ANOVA)

Analysis of Variance (ANOVA) is a numerical method, as discussed in [32], that can be used to check the significance of a particular sample. It can be expressed as given below in Eq. (1), Where F denotes the ANOVA coefficient, Mean Square Treatment (MST) denotes the Mean sum of squares due to treatment, and Mean Square Error (MSE) denotes the Mean sum of squares due to error.

$$F = \frac{MST}{MSE} \quad (1)$$

b) Chi-Squared

The chi-squared method is used to find the significance of each feature. The higher value obtained, as discussed in [33], denotes that the feature has more importance. Observed (O_i) is the actual data in the dataset, and expected (E_i) is the expected values based on the null hypothesis. It is calculated by Eq. (2).

$$\chi^2 = \Sigma \frac{(O_i - E_i)^2}{E^i} \quad (2)$$

2) Wrapper Methods (WFS)

Wrapper methods use a prediction model to find the best features. It is highly computational, as discussed in [34], and also gives the best performance and accuracy compared to embedded and filter methods.

a) Sequential Forward Selection (SFS)

Here is the content of Subsection (Level 4). It uses a greedy search algorithm. It twitches from the empty set ($\emptyset$) and eventually results in a high objective function with a full set (Y), as shown in Eq. (3).

$$SFS = \{\emptyset\} \rightarrow \{Y\} \quad (3)$$

b) Sequential Backward Selection (SBS)

R is the reverse of SFS. The algorithm twitches from the full set and results in the empty set, as shown in Eq. (4).

$$SBS = \{Y\} \rightarrow \{\emptyset\} \quad (4)$$

3) Embedded methods

Embedded methods are the common method used that use feature selection embedded with certain models. In this paper, we have used Least Absolute Shrinkage and Selection Operator (LASSO) and Regularized Least Squares Regression (RIDGE) techniques.

a) Lasso(L)

Its main scope is to shrink parameters that have no value to the model and also add a penalty, as described in [35], to the sum of coefficients. The Lasso can be expressed as given below in Eq. (5), where N is the number of cases, P is the covariates, x_i is the input/actual value, y_i is the predicted output, m is the slope complexity, Z is the intercept also known as Bias, λ is regularisation penalty.

$$L = \frac{1}{N} \sum_{i=1}^{N} (y_i - (mx_i + z))^2 + \lambda \sum_{i=1}^{p} (mx_i + z) \quad (5)$$

b) Ridge(R)

This technique is also known as L2 regularisation. The LASSO and RIDGE can be distinguished by the fact that the LASSO technique converts the coefficients to zero while the Ridge does not do the same as discussed in [36]. It can be expressed as given below in Eq. (6).

$$R = \frac{1}{N} \sum_{i=1}^{N} (y_i - (mx_i + z))^2 + \lambda \sum_{i=1}^{p} (mx_i + z)^2 \quad (6)$$

4) Proposed ensemble method

To detect whether or not plants are infected with a disease, the suggested technique uses a combination of feature sets that are produced through FS algorithms and random forest, as shown in Fig. 2. In order to combine all of the different approaches, hard voting is done. All of the individual models that make up the voting ensemble will be sent to a pipeline, and the first task of this pipeline will be to add a feature selection method in parallel with all of the other methods. This method is intended to select a particular number of features based on the best subset of each method, and it will be followed by a random forest classifier model, which determines whether or not the plant is healthy. The characteristic that is chosen by the vast majority of FS models receives the greatest amount of weighting in the evaluation process. It discovers all of the available subspaces and then calculates the subspace that provides the best result for the subset value (K). The k values, also known as the best features, range from 1 to 15, and whichever feature subspace demonstrates greater accuracy and MSE is selected as the feature subspace value for the proposed approach. The k values are found in the range of 1 to 15. The process of determining the optimal feature subset size (k) involves evaluating various subset sizes and selecting the one that strikes a balance between accuracy and complexity. In this context, $k = 8$ was chosen as the best-performing subset after considering several criteria.

Firstly, a range of subset sizes was likely assessed, and their corresponding model performance metrics, such as accuracy and MSE, were closely monitored. Ask increased, the model's accuracy typically improved as it had access to more features. However, beyond a certain point, adding more features could lead to overfitting, where the model becomes too complex and starts fitting noise in the data, resulting in a decline in accuracy on unseen data. The decision to choose $k = 8$ as the best-performing subset likely resulted from observing that it strikes the right balance. It provides a sufficient number of features to capture the essential information required for accurate plant disease prediction while avoiding excessive complexity. This complexity-accuracy trade-off ensures that the model remains robust and generalises well to new, unseen data. Other considerations, such as computational efficiency and interpretability, may have also influenced the choice of $k = 8$ as the optimal feature subset size.

The proposed approach is a holistic Ensemble Feature Selection (EFS) strategy that revolves around the harmonious integration of six distinct and carefully chosen feature selection algorithms. These algorithms, ANOVA, Chi-Square, Sequential Feature Selection (SFS and SBS), Lasso, and Ridge, have been meticulously tailored to excel in the task of identifying feature subsets of paramount relevance. The methodology adopted is inherently iterative, systematically exploring various feature count possibilities for each of the six algorithms. These iterative journeys culminate in the formation of pipelines, skillfully uniting feature selection and classification into a seamless and unified workflow. These pipelines are then consolidated into a collective entity termed `models`. The dataset in question is judiciously segregated into training and testing components, introducing the central figure, the VotingClassifier. This classifier, governed by the principles of hard voting, undertakes the role of a conductor within the ensemble. Its duty is to harmonise the decisions of the individual models, guided by the majority consensus of the ensemble's members. What truly sets this

approach apart is its democratic stance on features selected by the various algorithms; rather than imposing explicit rankings or weights, it promotes a collaborative philosophy where all contributions are respected in the decision-making process. This unity-driven approach strives to enhance predictive capabilities by synergising the strengths of diverse feature selection methods and classifiers, thus amplifying the ensemble's potential to deliver outcomes that are both robust and exceptionally accurate. While this research lays the groundwork for EFS, it refrains from delving into advanced mechanisms for explicit ranking or weighting of feature importance. This deliberate restraint opens the door to exciting prospects for future advancements in this domain.

A well-presented results section coupled with a convincing discussion will definitely prove the novelty and importance. The Voting Classifier (VC) classifies the data based on FS and RF algorithms as defined by the following Eq. (7).

$$VC = \begin{bmatrix} [X_1(RF)], [X_2(RF)], [X_3(RF)], \\ [X_4(RF)], [X_5(RF)], [X_6(RF)] \end{bmatrix} \quad (7)$$

The FS algorithm is used to select the best feature subsets(k) ranging from 1 to 15, as in Eq. (8).

$$X_N = [f_1, f_2, \dots, f_m] \quad for\ k = 0, \dots, 15 \quad (8)$$

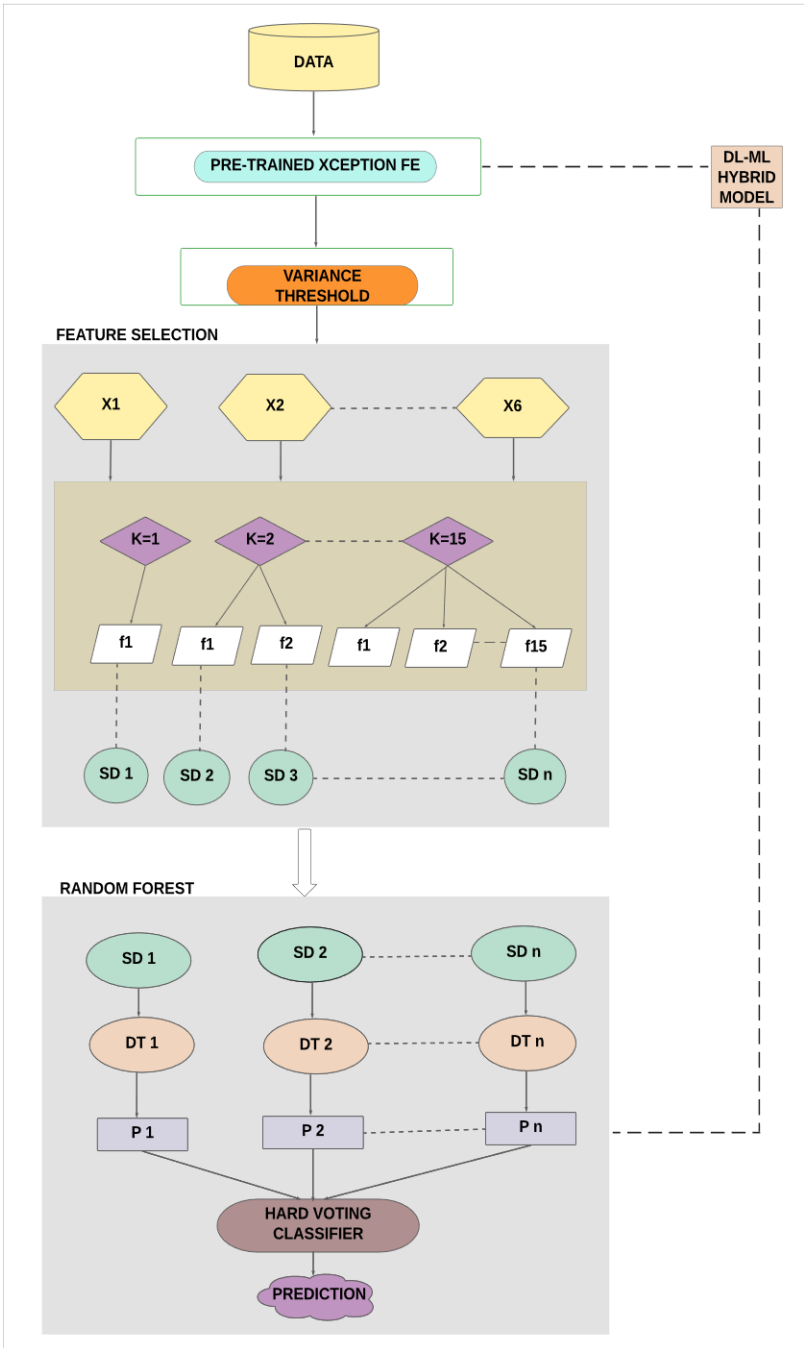

Fig. 2. System architecture of proposed method.

The RF first splits the data into subsets of data. Then, it performs Decision Tree (DT) classification in all the Subset Data (SD) as in Eq. (9).

$$RF = [SD_1(DT)], [SD_2(DT)], \ldots . [SD_n(DT)] \quad (9)$$

Finally, hard voting is done using the results obtained from Eqs. (8) and (9). The hard voting (*Y*) is calculated using Eq. (10).

$$Y = argmax \sum_{k=1}^{15} X_N(RF) \quad (10)$$

a) Algorithm

Algorithm 1. Ensemble Feature Selection (EFS)

Input data sets: Training datasets D1, D2
Feature Extraction: Features extracted from the Xception mode l(FE)
Feature Selection methods: ANOVA(X1), Chi-square(X2), SFS(X3), SBS(X4), Lasso(X5), Ridge(X6)
Classifier: Random Forest (RF)
Output: To find K best feature sets ranging from 1–15.
Step 1: Perform image pre-processing
Step 2: Extract features using the Xception model
Step 3: Apply the variance threshold method to remove the same subset features
Step 4: For D1, Initialize k to empty set $K = \{\emptyset\}$.
Step 5: While k=1, perform steps 6–10.
Step 6: Append FS methods with RF classifier for voting classifier as in Eq. (7).
Step 7: Perform feature selection of FS methods as in Eq. (8).
Feature Ranks=[F1 to Fi]
Step 8: Classify each FS method using the RF classifier as in Eq. (9).
Step 9: Perform hard voting on the classification results as in Eq. (10).
Step 10: Find the accuracy of the voting classification results for testing data as in Eq. (11).
Step 11: Repeat steps 6–10 for k = 2 to 15.
Step 12: Accuracy is obtained for k = 1 to 15.
Step 13: If $ki>kj$, Replace ki with new solution K.
Step 14: Find the best performing K value from step 10 using the accuracy.
Step 15: Postprocess the results and visualise.
Step 16: For D2, repeat steps 5 to 15.

b) Evaluation metrics

Dalianis [37] explored various evaluation criteria and provides valuable insights for a more transparent perspective when analyzing the results. In the quadrants, we have True Positives (TP), False Positives (FP), False Negatives (FN), and True Negatives (TN). Predictive accuracy can be expressed as the proportion of correctly classified results with true values by all other values as expressed below in Eq. (11). Mean squared error (MSE) is used to evaluate the quality of the prediction algorithms. It finds the sum of observed and predicted values. In Eq. (12), n stands for the number of data points, and the difference among y_i stands for the difference between observed and predicted values.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \quad (11)$$

$$MSE = \frac{1}{n} \sum_{1=1}^{n} (y_i - \hat{y}_i)^2 \quad (12)$$

IV. Result and Discussion

Experiments were run using each of the methods and almost every conceivable combination of variables. Each technique displays a distinct set of values in accordance with the feature sets that are chosen by the FS methods. The accuracy scores discovered from separate techniques of feature selection method of the dataset I, as shown in Table II, demonstrate performance that is comparable to that of the dataset II, as shown in Table III. The tables make it quite evident that RF has achieved superior results than those of other algorithms. The computed MSE values for dataset I can be found in Table IV, whereas the calculated MSE values for dataset II can be found in Table V. As shown in both tables, RF has shown a lower MSE when compared to other classifiers.

The implementation took place in a Jupyter Notebook on a Windows 11 Operating System S with 8 Giga Bytes of Random-access memory and an Intel i5 processor. As for hyperparameters, for feature selection methods like ANOVA and chi-square, the k value is an essential setting. Since there are fifteen features for the feature selection techniques, the suggested ensemble methods with the subset value k ranging from 1 to 15 are identified. For SFS, the hyperparameters are forward is set to True, and floating is set to False. For SBS, forward= False, floating = False. For lasso and Ridge, the alpha value is set to 1. The hyperparameters for LR and NB were set to Default. For SVM, the kernel was set to linear. For KNN n_neighbors were set to 7. For KNN, the random_state was set to 0. The hyperparameters used for the random forest classifier were with Random_state as 56 and n_estimators as 100. It is clear from the data shown in Tables II–V that RF performs better than other classifiers when it comes to identifying the state of our plants' health. As a result, the proposed ensemble technique makes use of RF for classification, drawing inferences from several individual implementations. Accuracy and mean squared error for the best feature subspace, K, are shown in Table V.

TABLE II. Accuracy Obtained from Dataset I

ML algorithms	Accuracy of Feature Selection Methods (%)					
	ANOVA	Chi-Squared	SFS	SBS	Lasso	Ridge
LR	81.0%	79.0%	87.0%	85.7%	81.7%	73.75%
NB	78.0%	74.5%	77.7%	76.2%	78.5%	72.75%
SVM	83.7%	83.5%	85.2%	86.7%	81.0%	73.0%
KNN	81.5%	61. %	58.7%	90.0%	72.0%	72.5%
CART	91.5%	91.2%	91.0%	92. %	88.7%	77.5%
RF	**95.2%**	**94.7%**	**95.7%**	**95.7%**	**94.0%**	**83.0%**

In Table II above, the Random Forest (RF) algorithm has demonstrated superior accuracy compared to other Machine Learning (ML) algorithms across all Feature Selection (FS) methods in Dataset I. NB and Ridge seem to have less accuracy below 80% compared to other ML and FS methods.

In Table III, the RF algorithm has performed better in terms of accuracy compared to the other ML algorithms across all the FS methods in Dataset II. It is visible that no matter the dataset, the algorithm stands firm in its performance with the best accuracy. In this table, all the algorithms perform relatively well in FS methods.

TABLE III. ACCURACY OBTAINED FROM DATASET II

ML algorithms	Accuracy of Feature Selection Methods (%)					
	ANOVA	Chi-Squared	SFS	SBS	Lasso	Ridge
LR	84.7%	81.2%	75.0%	87.5%	84.1%	82.5%
NB	80.0%	80.6%	76.8%	81.6%	80.6%	79.7%
SVM	82.2%	83.7%	78.1%	89.4%	84.4%	83.4%
KNN	74.7%	76.2%	74.0%	84.7%	82.5%	83.4%
CART	91.2%	93.1%	91.6%	92.9%	90.3%	87.5%
RF	**94.7%**	**95.9%**	**95.3%**	**95.6%**	**93.1%**	**88.75%**

In Table IV, the RF algorithm has performed better, as highlighted in the table, in terms of MSE, compared to the other ML algorithms across all the FS methods in Dataset I. In the above findings, NB and Ridge have performed the lowest.

In Table V, it is clear that the Random Forest (RF) algorithm outperforms other Machine Learning (ML) algorithms in terms of MSE across all Feature Selection (FS) methods for Dataset II. This demonstrates that RF consistently excels in predicting plant disease across various image datasets, as indicated by both accuracy and MSE metrics. it also reveals the lowest performing algorithms in dataset I, NB and Ridge, have performed relatively well in dataset II, one of the reasons it was included in the ensemble.

TABLE IV. THE MEAN SQUARE ERROR OF DATASET I

ML algorithms	Mean Squared Error of Feature Selection Methods (%)					
	ANOVA	Chi-Squared	SFS	SBS	Lasso	Ridge
LR	0.1675%	0.21%	0.13%	0.1425%	0.1825%	0.2625%
NB	0.22%	0.255%	0.2225%	0.2375%	0.215	0.2725%
SVM	0.1625%	0.165%	0.1475%	0.1325%	0.19%	0.27%
KNN	0.185%	0.3875%	0.4125%	0.1%	0.28%	0.275%
CART	0.085%	0.0875%	0.09%	0.0775%	0.1125%	0.225%
RF	**0.0475%**	**0.0525%**	**0.0425%**	**0.0425%**	**0.06%**	**0.17%**

TABLE V. MEAN SQUARE ERROR OF DATASET II

ML algorithms	Mean Squared Error of Feature Selection Methods (%)					
	ANOVA	Chi-Squared	SFS	SBS	Lasso	Ridge
LR	0.1531%	0.1875%	0.25%	0.125%	0.159375%	0.175%
NB	0.2%	0.1937%	0.2312%	0.1844%	0.1937%	0.2031%
SVM	0.1781%	0.1625%	0.2187%	0.1062%	0.1562%	0.1656%
KNN	0.2531%	0.2375%	0.2593%	0.1531%	0.175%	0.16562%
CART	0.0875%	0.06875%	0.0844%	0.0719%	0.0969%	0.125%
RF	**0.0531%**	**0.040625%**	**0.0469%**	**0.0437%**	**0.075%**	**0.1125%**

It is clear that setting *k* to 8 produces the most performing subset of features as in Table VI. Due to the fact that the performance of the two datasets using the ensemble technique reveals comparable findings, it is possible that the method that was suggested would be successful independently of any modifications. The comparison of the k values in the two datasets is shown in Figs. 3 and 4, respectively. The accuracy is evaluated, and a comparison of its performance with that of the current individual feature selection techniques in datasets I and II is carried out, as can be shown in Fig. 5. The MSE is used in the calculation of the error rate of the algorithms, as can be seen in Fig. 6.

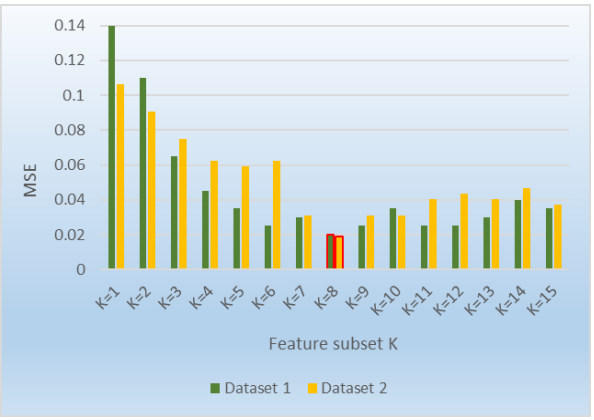

Fig. 4. Comparison of deciding the best subset value in terms of MSE.

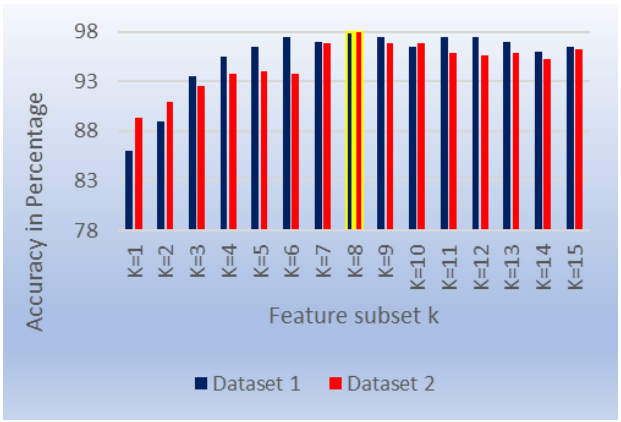

Fig. 3. Comparison of deciding the best subset value in terms of accuracy.

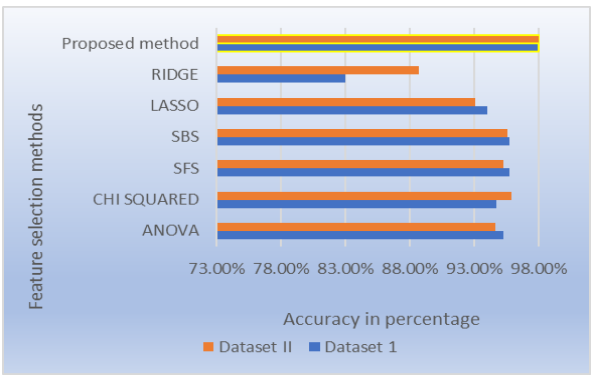

Fig. 5. Comparison in terms of accuracy for the proposed ensemble.

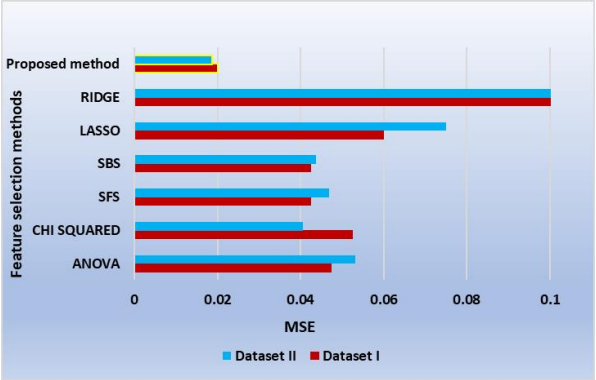

Fig. 6. Comparison in terms of MSE for the proposed ensemble.

TABLE VI. ACCURACY AND MSE OF K FEATURE SUBSPACE

Subset	Accuracy		MSE	
	Dataset I	Dataset II	Dataset I	Dataset II
K=1	86	89.375	0.14	0.10625
K=2	89	90.9375	0.11	0.090625
K=3	93.5	92.5	0.065	0.075
K=4	95.5	93.75	0.045	0.0625
K=5	96.5	94.0625	0.035	0.059375
K=6	97.5	93.75	0.025	0.0625
K=7	97	96.875	0.03	0.03125
K=8	**98**	**98.125**	**0.02**	**0.01875**
K=9	97.5	96.875	0.025	0.03125
K=10	96.5	96.875	0.035	0.03125
K=11	97.5	95.9375	0.025	0.040625
K=12	97.5	95.625	0.025	0.04375
K=13	97	959.375	0.03	0.040625
K=14	96	953.125	0.04	0.046875
K=15	96.5	96.25	0.035	0.0375

The study compares several feature selection methods, including ANOVA, chi-square, SFS, SBS, Lasso, and Ridge, with the proposed ensemble method, which combines all these techniques. ANOVA and chi-square are effective for categorical data, but they may struggle with complex relationships and continuous features. SFS and SBS offer flexibility in selecting subsets but can be computationally intensive. Lasso and Ridge mitigate overfitting but may not handle highly relevant features well. In contrast, the ensemble method harnesses the strengths of each technique to create a more robust feature subset. It aims to capture a broad range of feature relationships, both linear and nonlinear and offers improved predictive accuracy while accommodating various feature types and data distributions. Its performance should be empirically validated depending on the specific dataset and problem context. The differences observed in accuracy and MSE among the feature selection methods and algorithms can be attributed to several factors. These include variations in how the methods prioritise feature relevance, overfitting mitigation strategies, the synergy achieved through ensemble methods, the dataset's specific characteristics, such as noise and outliers, and the inherent sensitivity of machine learning algorithms to noisy or irrelevant features. For instance, ANOVA and chi-square might excel in selecting relevant categorical features but potentially overlook important continuous ones, while Lasso and Ridge regularisation techniques could be more adept at handling noisy data by penalising large coefficients. The proposed ensemble method combines the strengths of multiple feature selection techniques to enhance predictive accuracy and reduce MSE. The Xception-EFS-RF hybrid approach exhibits promise in plant disease prediction but comes with limitations. It may struggle to generalise effectively to new diseases or plant species not well-represented in the training data, potentially impacting accuracy in novel scenarios. Additionally, the computational demands for feature extraction and selection processes can be substantial, limiting scalability in resource-constrained environments. Researchers and practitioners should consider these limitations when applying Xception-EFS-RF in diverse contexts, as it is not tested in those fields.

The results presented in the tables above illustrate several notable trends and variations that shed light on the superior performance of the proposed method compared to other classifiers and feature selection techniques. The ensemble approach's strength lies in its aggregation of insights from multiple individual implementations. By harnessing RF's capabilities and mitigating the risk of overfitting, this approach results in a more robust and generalisable classification model, suitable for real-world agricultural applications. Optimal hyperparameter settings, such as the "*k*" values for ANOVA and chi-square, play a pivotal role in determining the quality of the feature subset. Fine-tuning these hyperparameters is essential to achieving the best results. Crucially, this research reflects the real-world applicability of the proposed approach by evaluating its performance on diverse datasets. This consideration of real-world agricultural scenarios underscores the method's practical value for real-time plant health assessment and crop management. In summary, the proposed ensemble approach's success can be attributed to the consistent performance of the RF classifier, the diverse feature selection techniques, and the robustness introduced by the ensemble approach. Careful hyperparameter tuning and the method's applicability to real-world scenarios contribute to its superior performance compared to other approaches.

V. CONCLUSION

Thus, an efficient method which could benefit agriculture and predict plant health is experimented with and evaluated. The use of CNN CNN-based Xception pre-trained model has extracted all the possible features for effective prediction. All the FS methods have performed quite well with the random forest algorithm. As the random forest is an ensemble algorithm, using an ensemble of FS method has been introduced along with it to enhance the performance of the model. The subset value (K) value is found in order to know which feature subset provides the best accuracy. When *K* is set to 8, it is found to be efficient compared to other feature subsets from 1 to 15. The proposed method has outperformed other algorithms in terms of accuracy and MSE. This study contributes a robust ensemble approach for plant disease prediction and classification. By integrating various feature selection techniques with the Random Forest (RF) classifier, the research consistently demonstrates superior performance

in accurately identifying plant health status. This method's versatility, with the inclusion of ANOVA, chi-square, SFS, SBS, Lasso, and Ridge, provides a well-rounded approach to feature selection, enhancing classification accuracy. The proposed ensemble approach holds significant practical implications for real-world agriculture, aiding in timely disease detection and crop management. Furthermore, the research emphasises the importance of hyperparameter optimisation, ensuring the best feature subsets are selected for optimal performance. In future, a well-defined booster optimisation method could be proposed using the above experiment results. Potential areas of research that can expand upon the above study include exploring different ensemble techniques, such as Random Forests and Stacking, to enhance predictive performance. Investigate the application of deep learning models, like CNNs and RNNs, for plant disease detection from image data and compare them with traditional machine learning models. Additionally, consider advanced image pre-processing methods, multimodal data fusion, and transfer learning across various plant species and diseases The future of real-time monitoring in agriculture lies in a comprehensive approach that integrates advanced technologies like the Internet of Things (IoT), Artificial Intelligence (AI), drones, cloud computing, and mobile applications. This visionary system, aligned with the principles of Agriculture 4.0, envisions a network of IoT sensors deployed in fields to collect real-time environmental data. AI algorithms process this data to detect early signs of plant diseases. Drones with high-resolution cameras capture aerial images for further analysis. Cloud-based platforms store and process the data, making it accessible to farmers through user-friendly mobile applications. The data collected through the IOT sensors could be stored in the Hadoop Distributed File System (HDFS) if the data generated is in large amounts. This visionary approach empowers farmers to make informed decisions and mitigate crop losses effectively.

Availability of Data and Materials

The data and materials will be made available upon reasonable request.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Abisha A: Collecting datasets, research concept and methodology, writing- original draft preparation. Bharathi N: Collecting datasets, reviewing, editing and Validation. All authors contributed to the article and approved the final version.

References

[1] A. Abisha and N. Bharathi, "Review on plant health and stress with various AI techniques and big data," in *Proc. 2021 International Conference on System, Computation, Automation and Networking (ICSCAN)*, 2021.

[2] S. Chakraborty and A. C. Newton, "Climate change, plant diseases and food security: An overview," *Plant Pathology*, vol. 60, no. 1, pp. 2–14, 2011.

[3] M. K. Choudhary and S. Hiranwal, "Feature selection algorithms for plant leaf classification: A survey," in *Proc. International Conference on Communication and Computational Technologies*, 2021, pp. 657–669.

[4] S. Aich *et al*., "A supervised machine learning approach using different feature selection techniques on voice datasets for prediction of Parkinson's disease," in *Proc. 2019 21st International Conference on Advanced Communication Technology*, 2019, pp. 1116–1121.

[5] G. Cisotto *et al*., "Feature selection for gesture recognition in internet-of-things for healthcare," in *Proc. 2020 IEEE International Conference on Communications (ICC 2020)*, 2020.

[6] M. Hashemi and M. Hall, "Visualisation, feature selection, machine learning: Identifying the responsible group for extreme acts of violence," *IEEE Access*, vol. 6, pp. 70164–70171, 2018.

[7] S. Y. Jiang and L. H. Wang, "Enhanced machine learning feature selection algorithm for cardiac arrhythmia in a personal healthcare application," in *Proc. 2018 IEEE Asia Pacific Conference on Postgraduate Research in Microelectronics and Electronics (PrimeAsia)*, 2018, pp. 39–42.

[8] O. Rado *et al.*, "Performance analysis of feature selection methods for classification of healthcare datasets," in *Proc. Intelligent Computing Conference*, 2019, pp. 929–938.

[9] T. Saw and P. H. Myint, "Feature selection to classify healthcare data using wrapper method with PSO search," *Int. J. Inf. Technol. Comput. Sci.*, vol. 11, no. 9, pp. 31–37, 2018.

[10] J. Ca *et al*., "Feature selection in machine learning: A new perspective," *Neurocomputing*, vol. 300, pp.70–79, 2018.

[11] B. Xue *et al*., "A survey on evolutionary computation approaches to feature selection," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 4, pp. 606–626, 2015.

[12] Z. Ceylan and A. Atalan, "Estimation of healthcare expenditure per capita of Turkey using artificial intelligence techniques with genetic algorithm-based feature selection," *Journal of Forecasting*, vol. 40, no. 2, pp. 279–290, 2021.

[13] R. Spencer *et al*., "Exploring feature selection and classification methods for predicting heart disease," *Digital Health*, vol. 6, 2020.

[14] J. Yu, "A hybrid feature selection scheme and self-organising map model for machine health assessment," *Applied Soft Computing*, vol. 11, no. 5, pp. 4041–4054, 2011.

[15] K. L. Chiew *et al.*, "A new hybrid ensemble feature selection framework for machine learning-based phishing detection system," *Information Sciences*, vol. 484, pp. 153–166, 2019.

[16] P. I. Metselaar *et al.*, "Recursive ensemble feature selection provides a robust mRNA expression signature for myalgic encephalomyelitis/chronic fatigue syndrome," *Scientific Reports*, vol.11, no. 1, pp. 1–11, 2021.

[17] A. Hashemi *et al.*, "A pareto-based ensemble of feature selection algorithms," *Expert Systems with Applications*, vol. 180, 115130, 2021.

[18] G. Yao, X. Hu, and G. Wang, "A novel ensemble feature selection method by integrating multiple ranking information combined with an SVM ensemble model for enterprise credit risk prediction in the supply chain," *Expert Systems with Applications*, vol. 200, 117002, 2022.

[19] B. S. Pardo *et al*., "Ensemble feature selection: Homogeneous and heterogeneous approaches," *Knowledge-Based Systems*, vol. 118, pp.124–139, 2017.

[20] K. Makimoto *et al*., "Comparison of feature selection methods and machine learning classifiers for predicting chronic obstructive pulmonary disease using texture-based CT Lung radiomic features," *Academic Radiology*, vol. 30, 2022.

[21] D. Aqel *et al.*, "Extreme learning machine for plant diseases classification: A sustainable approach for smart agriculture," *Cluster Computing*, vol. 25, no. 3, pp. 2007–2020, 2022.

[22] A. Hashemi *et al*., "Ensemble of feature selection algorithms: A multi-criteria decision-making approach," *International Journal of Machine Learning and Cybernetics*, vol. 13, no. 1, pp. 49–69, 2022.

[23] A. Y. Notash *et al*., "Evolutionary ensemble feature selection learning for image-based assessment of lymphedema arm volume," *Concurrency and Computation: Practice and Experience*, vol. 34, no. 1, 2022.

[24] J. L. Leevy *et al*., "IoT information theft prediction using ensemble feature selection," *Journal of Big Data*, vol. 9, no. 1, pp. 1–48, 2022.
[25] D. Hughes and M. Salathé, "An open-access repository of images on plant health to enable the development of mobile disease diagnostics," arXiv preprint, arXiv:1511, 2015.
[26] E. Medhi and N. Deb, "PSFD-Musa: A dataset of banana plant, stem, fruit, leaf, and disease," *Data in Brief*, vol. 43, 108427, 2022.
[27] K. A. Mahmud. (2021). Banana Leaf Dataset. [Online]. Available: https://www.kaggle.com/datasets/kaiesalmahmud/banana-leaf-dataset
[28] V. K. Vishnoi *et al*., "Plant disease detection using computational intelligence and image processing," *Journal of Plant Diseases and Protection*, vol. 128, no. 1, pp. 19–53, 2021.
[29] A. Abisha and N. Bharathi, "Feature extraction from plant leaves and classification of plant health using machine learning," *Advanced Machine Intelligence and Signal Processing*, pp. 867–876, 2022.
[30] A. Abisha and N. Bharathi, "A hybrid feature extraction and classification using Xception-RF for multiclass disease classification in plant leaves," *Artificial Intelligence*, vol. 37, no. 1, 2023.
[31] N. Nandhini and R. Bhavani, "Feature extraction for diseased leaf image classification using machine learning," in *Proc. 2020 International Conference on Computer Communication and Informatics (ICCCI)*, 2020.
[32] L. M. Connelly, "Introduction to Analysis of Variance (ANOVA)," *Medsurg Nursing*, vol. 30, no. 3, pp. 218–158, 2021.
[33] N. S. Turhan, "Karl Pearson's chi-square tests," *Educational Research and Reviews*, vol. 16, no. 9, pp. 575–580, 2020.
[34] N. E. Aboudi and L. Benhlima, "Review on wrapper feature selection approaches," in *Proc. 2016 International Conference on Engineering and MIS (ICEMIS)*, 2016.
[35] R. Muthukrishnan and R. Rohini, "LASSO: A feature selection technique in predictive modeling for machine learning," in *Proc. 2016 IEEE International Conference on Advances in Computer Applications (ICACA)*, 2016, pp. 18–20.
[36] P. Zhang *et al*., "A feature selection method combined with ridge regression and recursive feature elimination in quantitative analysis of laser-induced breakdown spectroscopy," *Plasma Science and Technology*, vol. 22, no. 7, 2020.
[37] H. Dalianis, "Evaluation metrics and evaluation," *Clinical Text Mining*, pp. 45–53, 2018.