A SYNERGISTIC FEATURE SELECTION FRAMEWORK INTEGRATING STATISTICAL TESTS AND SEQUENTIAL SELECTION FOR IMPROVED PLANT DISEASE DIAGNOSIS

Khudaiberdiev M.Kh, Madrakhimov A.Kh, Muraeva Kh.M

Department of Software of Information Technologies, Tashkent University of information technologies named after Muhammad Al-Khwarizmi, Tashkent, Uzbekistan

ABSTRACT

Achieving high predictive accuracy while utilizing a minimal yet highly relevant set of features remains a critical challenge in machine learning. To address this, numerous feature selection techniques—ranging from ANOVA and LASSO to supervised methods like Mutual Information—have been developed, each offering unique strengths and limitations. Building upon these foundations, we introduce innovative hybrid approaches that seamlessly integrate filter and wrapper methods for more effective feature selection. These approaches were rigorously tested across multiple models. In this research, the new approaches of the integration of Chi-square test (with new step: relationship level) with SFS algorithm are proposed for enhancement model performance in feature selection. By the help of the new approaches, the trained results were 89% (in SVM), 88% (in kNN), 88% (in RF), 91% (in FCNN). These outputs are better than other feature selection methods' results.

KEYWORDS

Feature Selection, machine learning, dimensionality, Chi-square test, Sequential Forward Selection, classification, statistical features.

1. INTRODUCTION

Early and accurate detection of plant leaf diseases is critical for reducing crop losses and ensuring sustainable agricultural productivity. With the advancement of computer vision and machine learning techniques, image-based disease diagnosis has become a prominent research direction. In recent years, a variety of feature extraction and classification algorithms have been proposed to identify and distinguish between different plant diseases based on leaf characteristics.

The rapid increase in data dimensionality creates significant challenges for most mining and learning algorithms, including issues like the curse of dimensionality, high storage demands, and increased computational costs. Feature selection has proven to be an effective and efficient method for preparing high-dimensional data for data mining and machine learning tasks. By integrating cutting-edge techniques and diverse feature sets, the field of feature selection has not only progressed but also adapted over time, making it suitable for an increasingly wide spectrum of applications. This entry aims to provide a fundamental overview of feature selection, covering its basic concepts, classifications of current methods, recent advancements, and practical uses [1]. In order to handle high-dimensional datasets more effectively, data mining employs dimensionality reduction methods. Such methods generally focus on either deriving additional features or

DOI:10.5121/acij.2025.16501

identifying the most significant attributes from the existing feature set. [2]. Feature selection is primarily employed to reduce computational cost and memory usage, mitigate the effects of the curse of dimensionality, and lower the risk of overfitting. These improvements contribute to better generalization and enhance the overall performance of machine learning models [3].

In this study, we propose a hybrid approach to achieve optimal model accuracy with minimal features, outperforming conventional methods (Chi-square test, Mutual Information, SFS, etc.). This approach innovatively integrates Cramér's V coefficient into the Chi-square test algorithm and synergistically combines the output with the Sequential Forward Selection algorithm. Unlike methods like SHAP or Permutation Importance that are model-dependent and computationally expensive, our approach allows for direct interpretation of a feature's usefulness in discriminating between disease classes using already trained feature subsets. The proposed method is tested on a tomato leaf disease dataset with ten classes and shows promising results in identifying class-discriminative features while maintaining computational efficiency.

The rest of the paper is structured as follows: Section 2 discusses the materials and preprocessing steps. Section 3 describes the proposed methodology in detail. Section 4 presents the experimental results and analysis. Finally, Section 5 concludes the paper and outlines future directions.

2. RELATED WORKS

In recent years, numerous studies have emphasized the importance of feature selection for accurate plant disease classification. High-dimensional image data, especially from leaf images, often contain redundant or irrelevant features that may hinder classification performance. To overcome this, various statistical and machine learning-based feature selection methods have been proposed. One notable study utilized six color features and twenty-two texture features, including metrics derived from the Gray-Level Co-occurrence Matrix (GLCM), to characterize diseaseaffected plant leaves. To identify the most discriminative features, the authors[9] applied Chisquare test and ANOVA (Analysis of Variance), which are widely used statistical methods for feature ranking based on class separability. These methods allowed the researchers to reduce the dimensionality of the feature set while retaining its discriminative power.This demonstrates the usefulness of statistical filtering methods in reducing dimensionality and retaining only the most discriminative features, thereby improving overall model efficiency.

In the study [10] conducted by Nisar Ahmad and colleagues, 6 color and 22 texture (GLCM-based) features were extracted, and Chi-square and ANOVA tests were used to select the most informative ones. This approach, implemented using the SVM model with 10-fold crossvalidation, demonstrating its effectiveness among feature-based methods. The paper highlights that filtering with the Chi-square test helped eliminate redundant features, thereby improving both the speed and reliability of subsequent models.

1.1. Feature Extraction and Feature Selection

The rapid expansion of high-dimensional data on the internet in recent years has posed major difficulties for machine learning algorithms, especially in managing extensive feature sets. To address these challenges, preprocessing steps have become essential to ensure the effective application of machine learning techniques. Among these, feature selection plays a crucial role as it helps reduce dimensionality and enhances the overall performance of learning algorithms[1]. Specifically, in tomato disease detection tasks, selecting optimal features helps reduce computational complexity, improve model interpretability, and increase classification accuracy. By eliminating noisy or redundant features, the model becomes more robust and less prone to

overfitting, particularly in cases where inter-class variability is subtle. Furthermore, feature selection contributes to building lightweight and efficient diagnostic systems suitable for real-time or resource-constrained agricultural applications.

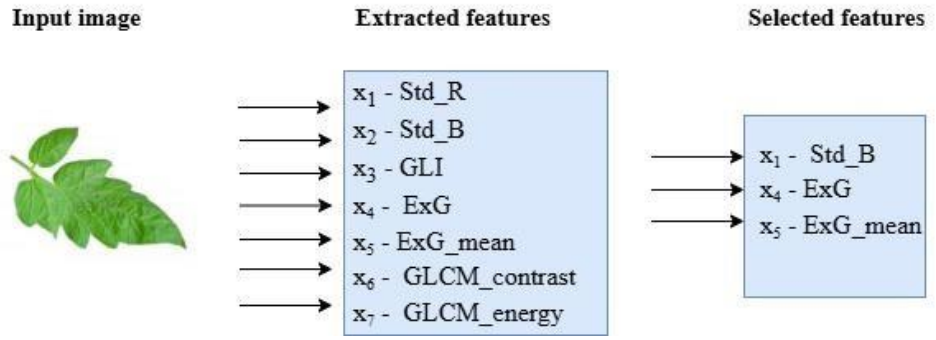

Figure 1. The process of feature selection from tomato leaf images

Feature selection is divided into three major approaches: Filter, Wrapper and Embedded approach.

Filters. Filters are feature selection methods that do not use any modeling algorithms. Instead of outside information, they focus on how features in the training data are related to the target class. [2]. Compared to Wrappers, filter methods are faster and tend to generalize better.

Wrappers. Wrappers assess features by considering their interactions with each other, unlike filters that evaluate features individually. These methods evaluate and compare different combinations of features using learning algorithms, which can make training slower and more complicated.[3].

Embedded. Embedded methods perform feature selection and optimization simultaneously during the classification process. This approach can identify feature dependencies with less computational complexity compared to wrapper methods [4].

3. METHODOLOGY

In this research work, the dataset “Tomato disease ” is used which is taken from the PlantVillage in the Kaggle.com. Bilateral filtering was employed for noise reduction, while CLAHE (Contrast Limited Adaptive Histogram Equalization) was applied to enhance image contrast during the preprocessing stage.

Figure 1. On the left side: original image, On the right side: After using of bilateral filtering method.

In scientific research, to extract the color, texture, shape, and statistical characteristics from plant leaf images, an ensemble of algorithms was created to extract the features of each category. Suppose we have a training set M. A feature set $X_i=\{x_{i1},x_{i2}, ...x_{nm}\}$ must be derived for every object S_i in dataset, where n denotes the total number of objects, m represents the number of features extracted for each object. The ensemble of algorithms for extracting features of different categories is organized as follows:

Step 1: Separate the image into RGB color channels.

Every pixel is represented by a combination of three color components: red, green, and blue. We analyze each channel separately.

Step 2: Calculate color statistics (color moments):

1.Mean (Average)

$$M_i = \sum_{j=1}^{N} \frac{1}{N} P_{ij}$$

P_{ij} is the value of the i -th color channel at the j -th pixel, N is the total number of pixels, M_i is the average color value in the -i th channel.

2. Standard Deviation. Measures the spread (dispersion) of color values around the mean in the i -th channel.

$$CD_i = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (P_{ij} - M_i)^2}$$

3. Skewness. Measures the asymmetry of the color distribution in the i -th channel.

$$SK_i = \sqrt[3]{\frac{1}{N} \sum_{j=1}^{N} (P_{ij} - M_i)^3}$$

4. Kurtosis. Kurtosis quantifies the shape characteristics of the color distribution in the -i th channel:

$$K_i = \sqrt[4]{\frac{1}{N} \sum_{j=1}^{N} (P_{ij} - M_i)^4}$$

Step 3: Calculating color indexes

Excess Green Index (ExG):

Used to quantify the dominance of green color in an image:

$$ExG = 2g - r - b$$

here

$$r = \frac{R^*}{(R^* + G^* + B^*)}, \quad g = \frac{G^*}{(R^* + G^* + B^*)}, \quad b = \frac{B^*}{(R^* + G^* + B^*)}$$

$$R^* = \frac{R}{R\max}, \quad G^* = \frac{G}{G\max}, \quad B^* = \frac{B}{B\max}$$

Green Leaf Index. GLI measures the greenness level of plant leaves using RGB values and produces a grayscale image with pixel values between -1 and 1.

$$GLI = \frac{2G - R - B}{2G + R + B}$$

Color Ratios:

These ratios measure the relative dominance of colors:

$$Red - Green - Ratio = \frac{R}{G}, \quad Blue - Green - Ratio = \frac{B}{G}, \quad Red - Blue - Ratio = \frac{R}{B}$$

Step 4. Calculating texture features.

1. Contrast. Contrast quantifies the variation in pixel intensities across an image. A high contrast means large intensity differences; low contrast means pixels have similar intensities.

$$Contrast = \sum_i \sum_j (i-j)^2 P(i,j)$$

2. *Correlation*. Correlation measures how correlated a pixel is to its neighbor over the whole image. High correlation means strong linear relationship between pixel pairs.

$$Correlation = \frac{\sum_i \sum_j [ij * P(i,j) - \mu_x * \mu_y]}{\sigma_x * \sigma_y};$$

where

i and j are GLCM indices.

$P(i,j)$ is the GLCM element.

In the Gray Level Co-occurrence Matrix (GLCM), μ_x and μ_y represent the mean values of the row and column distributions, respectively, while σ_x and σ_y denote their corresponding standard deviations.

3. *Energy*.Energy quantifies the degree of texture uniformity by calculating the sum of squared values in the Gray Level Co-occurrence Matrix (GLCM)..

$$Energy = \sum_i \sum_j P(i,j)^2$$

4. Homogeneity. Homogeneity measures how close the elements of the GLCM are to the diagonal, indicating similarity between neighboring pixel values.

$$Homogenity = \sum_i \sum_j \frac{P(i,j)}{1 + |i-j|}$$

Step 5: Adding the results to the feature table.

By using of feature extraction algorithms, total of 29 features are extracted. These features are belong to color, texture, shape and statistical features of tomato leaf images. Here some feauture extraction process is illustrated:

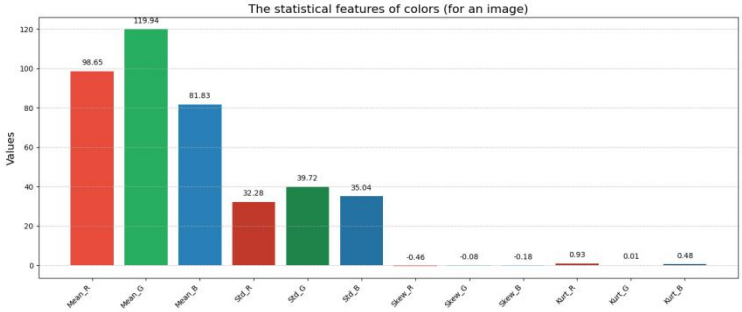

Figure 1. The statistical features of colors

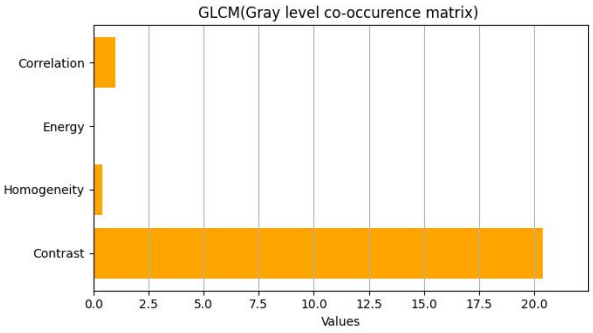

Figure 2. The texture features of an image

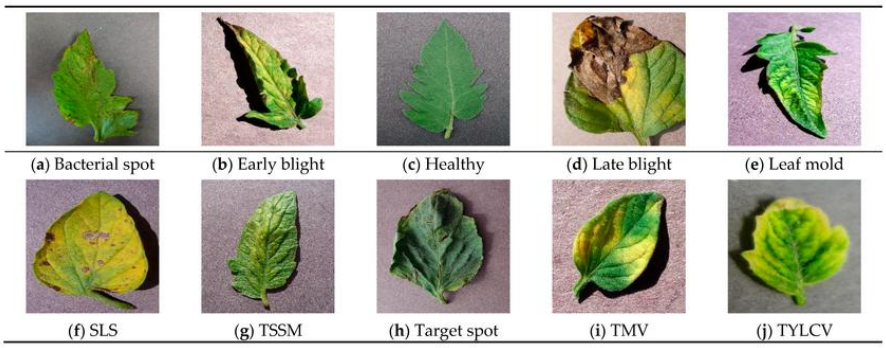

Figure 3. The samples in the dataset "Tomato disease"

There are given two algorithms that combine to select informative features from dataset. These two algorithm are related to each other. At first the Chi-test algorithm are implemented. We added new step to the algortihm which is named relationship level (based on Cramers'V). After that the set of output features were combined to the SFS(Sequential Forward Selection) method. The algorithms of the proposed approaches are following:

Algorithm 1. Chi-Square test with Cramer's V

Input: Contingency table with observed frequencies K_{ij}

Output: Association decision (Reject or Fail to reject H_0) a selected new feature set F

1. Defining hypotesis:

H_0 - null hypotesis, There is not relation between classes and features;

H_1 - alternative hypotesis, There is relation between classes and features

2. Compute expected frequencies:

For each cell (i, j) :

$$B_{ij} = (i\_rows\_total \times i\_columns\_total) / Grand\_total$$

3. Compute statistics φ :

$\varphi = 0$

For each cell (i, j) :

$$\varphi = \varphi + (K_{ij} - B_{ij})\text{^}2 / B_{ij}$$

4. Compute Cramér's V δ_{bd} :

$$\delta_{bd} = \sqrt{\varphi / N \times \min(rows\_total - 1, columns\_total - 1)}$$

5. Determine critical value:

$$df = (rows\_total - 1) \times (columns\_total - 1)$$

$\varphi_{critics} = \varphi$ taken from the table φ for critics value (α, df)

6. Compare:

If- $\varphi_{critics} > \varphi$

Reject H_0

Else:

Fail to reject H_0

7. Interpret association strength:

$0.0 < \delta_{bd} \le 0.1$: No association

$0.1 < \delta_{bd} \le 0.35$: Weak association

$0.35 < \delta_{bd} \le 0.5$: A moderately strong association

$\delta_{bd} > 0.5$: Strong association

8. Final set $F = \{f_1, f_2, ..., f_m\}$

Algorithm 2. Union of SFS+ Chi-Square test with Cramer's V

Input:

$F = \{f_1, f_2, ..., f_m\}$ // Features selected by correlation test

p // Number of features to select

Y // Full set of features

Initialize:

$X_0 = \varnothing$ // Start with empty selected feature set

$k = 0$ // Counter for selected features

Step 2: Sequential forward feature selection

while $k < p$ do:

For each x in $(Y - X_k)$:

Compute $J(X_k \bigcup \{x\})$ // Evaluate criterion function for adding x

End For

$X^+ = \arg\max\_ xJ(X_k \cup \{x\})$ // Select feature x that maximizes J

$X(k+1) = X_k \cup \{X+\}$ // Add selected feature to set

$k = k + 1$ // Increment counter

end while

// Step 4: Combine features selected by correlation and sequential selection

$$Z = F \cup X$$

// Step 5: Final set Z

Output:

U // Final combined feature set

4. RESULTS

The obtained results clearly indicate that the hybrid feature selection strategy, which combines statistical relevance (Chi-Square test) with sequential evaluation (SFS), leads to significant improvements in classification performance. This outcome demonstrates the strength of leveraging both global statistical importance and local feature interactions in the selection process. Notably, the use of color-based and texture-based features allowed for more precise differentiation among disease classes, as evidenced by the high F1-scores in the selected combinations. The analysis of feature importance on a per-class basis further provided interpretability into how individual features contributed to the classification of specific diseases. For instance, features like Mean_R and GLCM_correlation consistently appeared in the top-performing combinations, indicating their robustness and diagnostic relevance. Several machine learning algorithms were applied using the selected subset of features. Their performance metrics were systematically compared and summarized in a comparative evaluation table to highlight the effectiveness of each model.

Table 1. Comparative study

№	Methods / Algorithms	Support Vector Machine (accuracy)	k-Nearest Neighbors (accuracy)	Random Forest (accuracy)	Fully Connected Neural Network (accuracy)
1	Chi-test with Cramer's V (9 features)	85%	83%	86%	90%
2	Chi-test (16 features)	84%	81%	85%	89%
3	SFS(10 features)	85%	83%	84%	89%
4	Hybrid(13 features)	89%	87%	88%	91%
5	Mutual information (14 features)	84%	79%	85%	87%
6	ANOVA(9 features)	77%	77%	81%	77%
7	LASSO (12 features)	86%	84%	86%	89%
8	Full features (29 features)	86%	73%	87%	91%

5. CONCLUSION

In this study, we proposed novel hybrid feature selection methods that effectively combine filter and wrapper approaches, specifically integrating the Chi-square test with Cramér's V coefficient and Sequential Forward Selection. Our experimental results on the tomato disease dataset demonstrate that the hybrid approach outperforms conventional feature selection techniques such as ANOVA, LASSO, and Mutual Information in terms of classification accuracy across multiple machine learning models. The selection of fewer but more meaningful features allows the proposed method to reduce processing time and simultaneously improve the clarity and stability of the model.

These outcomes underline the effectiveness of hybrid feature selection in boosting the accuracy of machine learning systems for agricultural disease detection and suggest opportunities for expanding these methods to new domains using other selection metrics. Looking ahead, future research may focus on extending this approach to other agricultural datasets, including those involving different plant species or image modalities such as hyperspectral or temporal data. Furthermore, adaptive or dynamic feature selection mechanisms could be developed based on class-specific performance analysis. Incorporating expert domain knowledge and integrating additional statistical or modelagnostic criteria may also help to refine the feature selection process and improve applicability in real-world agricultural decision-support systems.

REFERENCES

[1] Suhang Wang, Jiliang Tang and Huan Liu, Feature selection, Springer Science+Business Media New York 2016.

[2] Akhiat, Y., Asnaoui, Y., Chahhou, M., & Zinedine, A. A new graph feature selection approach. In 2020 6th IEEE Congress on Information Science and Technology (CiSt) (pp. 156-161). IEEE. (2021, June).

[3] Akhiat, Y., Manzali, Y., Chahhou, M., & Zinedine, A. A New Noisy Random Forest Based Method for Feature Selection. Cybernetics and Information Technologies, 21(2), 10-28. (2021).

[4] Vipin Kumar and Sonajharia Minz. Multi-view ensemble learning: A supervised feature set partitioning for high dimensional data classification. In Proceedings of the Third International Symposium on Women in Computing and Informatics, pages 31–37. ACM, 2015.

[5] Tony Bellotti a, Ilia Nouretdinov b, Meng Yang b, Alexander Gammerman, Chapter 6 - Feature Selection, Conformal Prediction for Reliable Machine Learning, Theory, Adaptations and Applications, 2014, Pages 115-130

[6] Damodar Patel, Amit Saxena, John Wang, A Machine Learning-Based Wrapper Method for Feature Selection, International Journal of Data Warehousing and Mining Volume 20 • Issue 1. January-December 2024

[7] Jianuo Li, Hongyan Zhang, Jianjun Zhao, Xiaoyi Guo,Wu Rihan and Guorong Deng, Embedded Feature Selection and Machine Learning Methods for Flash Flood Susceptibility-Mapping in the Mainstream Songhua River Basin, China, . Remote Sens. 2022, 14, 5523.

[8] Younes Bouchlaghem and Yassine Akhiat and Souad Amjad, Feature Selection: A Review and Comparative Study, E3S Web of Conferences 351, 01046 (2022).

[9] Weinan Li, Lisen Liu, Jianing Li , Weiguang Yang, Yang Guo, Longyu Huang, Zhaoen Yang, Jun Peng, Xiuliang Jin and Yubin Lan, Spectroscopic detection of cotton Verticillium wilt by spectral feature selection and machine learning methods,Front. Plant Sci. 16:1519001. doi: 10.3389/fpls.2025.1519001(2025).

[10] Nisar Ahmed*, Hafiz Muhammad Shahzad Asif, Gulshan Saleem, Leaf Image based Plant Disease Identification using Color and Texture Features, Wireless Personal Communications, 2021, pages 12-14.