Advances in Computer, Signals and Systems (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/acss.2025.090207
ISSN 2371-8838 Vol. 9 Num. 2

Research on Wheat Seed Classification Based on Machine Learning Algorithms and Data Analysis Visualization

Kaili Sun[a,*], Wei Bai, Jiexin Feng, Zhe Yang, Yanyan Li

School of Trade and Economic, Haojing College of Shaanxi University of Science & Technology, Xi'an, Shaanxi, China
[a]18434760894@163.com
[*]Corresponding author

Keywords: Machine Learning, Data Analysis and Visualization, Feature Combination

Abstract: This study addresses the problem of wheat seed classification by employing three machine learning algorithms—Random Forest (RF), Naïve Bayes (NB), and Support Vector Machine (SVM)—on the Wheat Seeds Dataset from the UCI database. Through comprehensive data preprocessing, feature analysis, and model construction, the impact of different feature combinations on classification accuracy was systematically investigated. The dataset, comprising 210 samples with seven attributes (e.g., area, perimeter, and kernel groove length), was standardized and split into training and testing sets to ensure robust evaluation. The experimental results demonstrate that RF and SVM significantly outperform NB in classification performance, with SVM achieving the highest accuracy of 97.61% when combining area or width with kernel groove length. Notably, the combination of perimeter and kernel groove length yielded the highest accuracy (96.67%) in RF, while compactness and asymmetry coefficient consistently performed poorly across all algorithms, with accuracy as low as 60.71% in SVM. These findings highlight the critical role of feature selection in classification tasks, with kernel groove length emerging as a key determinant. This research not only provides an effective technical reference for wheat variety classification but also underscores the practical value of machine learning in agricultural applications, offering insights for optimizing efficiency and reducing costs in food security initiatives.

1. Introduction

Wheat is the most important global food crop, leading in cultivation area, production volume, and trade value. Its development directly impacts national food security and social stability. However, different wheat varieties exhibit variations in yield, economic benefits, and susceptibility to diseases. Thus, achieving high accuracy in wheat variety classification is a research topic of great significance.

The essence of machine learning lies in algorithms, technically defined as searching for useful representations of input data within a predefined possibility space, guided by feedback signals [1].

Applying machine learning methods to wheat seed classification can significantly improve processing efficiency and classification accuracy while reducing labor, material, and financial costs, making it highly practical [2].

This paper selects the wheat seed Data Set in the UCI database and divides it into the training set and the test set. Three algorithms in machine learning, namely Random Forest, Naive Bayes and Support Vector Machine (SVM), were selected to construct their models respectively to classify wheat seeds, and the influence of different feature combinations on the classification accuracy of machine learning methods was discussed [3]. Then, by comparing and analyzing the experimental results, the classification accuracy was selected as the evaluation index of the model to compare the three machine learning methods, and the feature combination with the best performance in exploring the classification of wheat varieties was obtained.

2. Data Preprocessing

The dataset used in this study is the "Seeds Data Set" from the UCI database, comprising 210 samples. Each sample includes seven measured attributes: area, perimeter, compactness, kernel length, kernel width, asymmetry coefficient, and kernel groove length. The compactness C is calculated as $C = \frac{4\pi A}{P^2}$.

2.1 Data Encoding

The wheat varieties are categorized into three classes: samples 1–70 are Kama (label 1), samples 71–140 are Rosa (label 2), and samples 141–210 are Canadian (label 3), as shown in Table 1.

Table 1: Wheat Variety Labels

Variety	Label
Kama	1
Rosa	2
Canadian	3

2.2 Data Standardization

Descriptive statistical analysis revealed inconsistencies in the scales of the attributes, which could affect the results. To address this, all 210 samples were standardized to the range [0, 1] using:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (1)$$

The dataset was shuffled and split into training (60%, 126 samples) and testing sets (40%, 84 samples) using the ‘train_test_split’ function from the scikit-learn library.

2.3 Data Correlation Visualization

Scatter plots, also known as X-Y plots, display all data points on a Cartesian coordinate system. By observing the distribution of data points on a scatter plot, the correlation between variables can be inferred. If there is no relationship between the variables, the data points will be randomly distributed and discrete on the scatter plot. If there is a certain correlation, most of the data points will be relatively concentrated and show a certain trend [4].

This article has drawn a scatter plot matrix, visualizing the distribution of features in the dataset, thereby examining the relationships between pairs of feature attributes. From Figure 1, it can be

observed that the area of wheat, its perimeter, grain length, grain width, and groove length show a clear linear relationship; while the relationship between the compactness and asymmetry coefficient of wheat and other features is more complex and difficult to represent with a linear relationship.

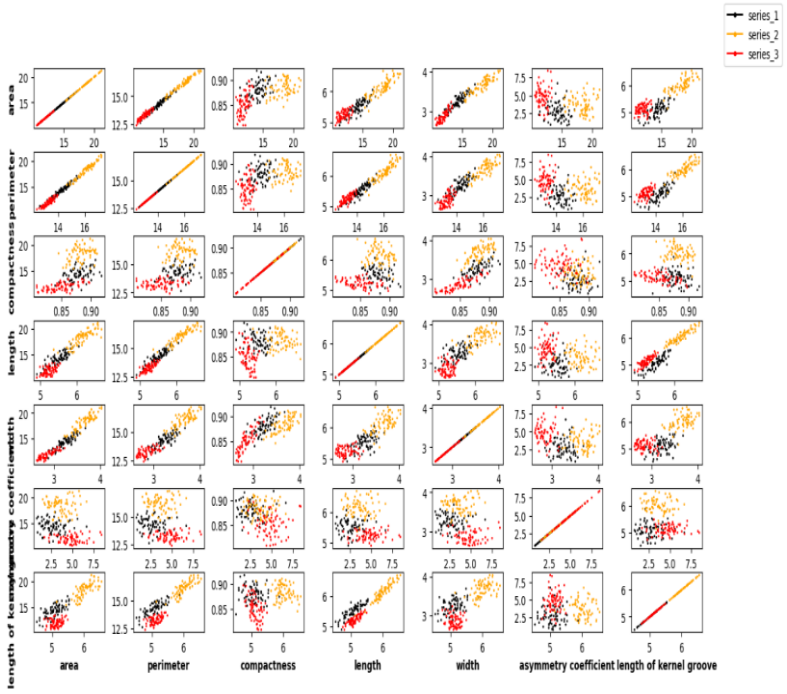

Figure 1: Scatter Matrix of the Seeds Dataset

3. Machine learning algorithms for classification of wheat datasets and data analysis visualization

3.1 Random Forest (RF)

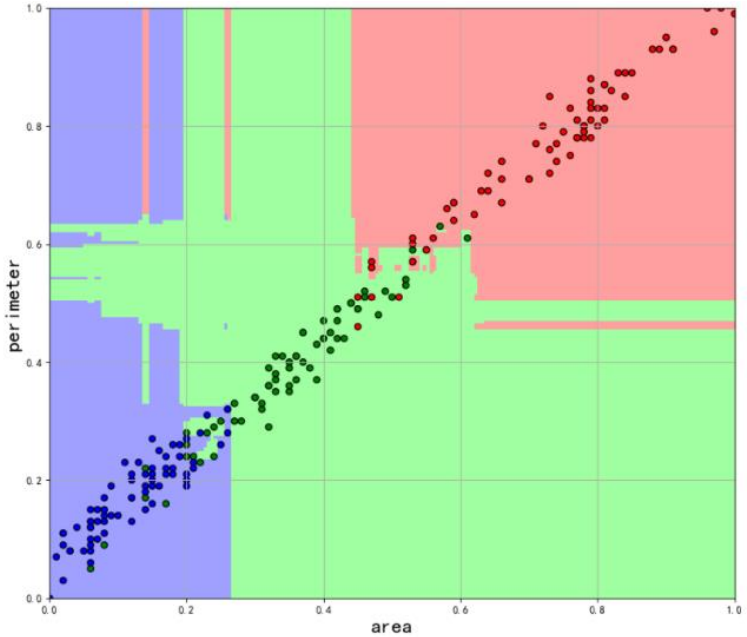

Figure 2: Classification Results for Area + Perimeter (RF)

Random Forest (RF) is a statistical learning theory. It uses the bootstrap resampling method to draw multiple samples from the original sample, builds decision tree models for each bootstrap sample, and then combines the predictions of multiple decision trees through voting to obtain the final prediction result. Moreover, Random Forest has a very high prediction accuracy, a good tolerance for outliers and noise, and is less likely to overfit [5].

Based on the above theory, we used the wheat seed dataset to explore the relationship between different pairs of attributes and the type of wheat. To avoid overfitting in decision trees and explore the correlation of attributes among different types of wheat, we constructed a Random Forest model. Figure 2 shows the classification results of area + perimeter.

The number of correct data predictions was 194, and its accuracy rate reached 92.38%, demonstrating a relatively high classification effect. Next, combine the seven attributes in pairs to obtain the prediction results shown in Figure 3:

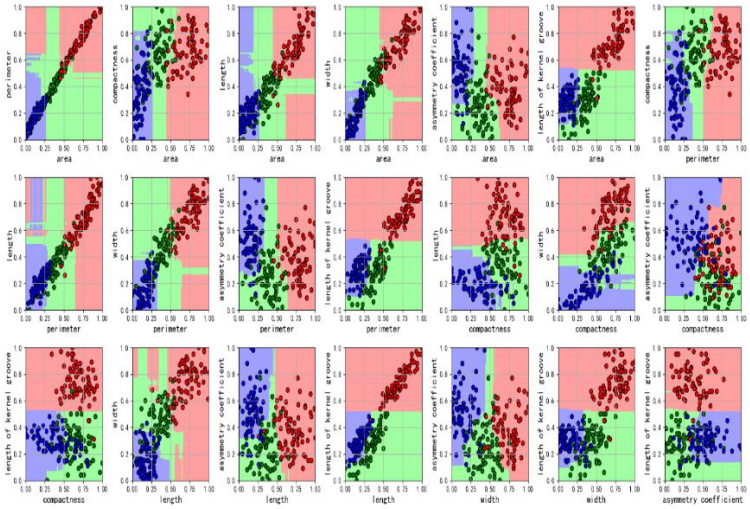

Figure 3: Impact of Feature Pairs on Classification (RF Algorithm)

The following prediction accuracy rates are obtained:

Random Forest (RF)							
Attributes	area	perimeter	compactness	length	width	asymmetry coefficient	length of kernel groove
area		92.38%	91.43%	90.00%	91.90%	94.29%	95.24%
perimeter			91.90%	90.00%	90.95%	95.71%	96.67%
compactness				90.00%	90.00%	72.38%	86.67%
length					89.05%	90.00%	94.76%
width						94.76%	92.86%
asymmetry coefficient							88.57%
length of kernel groove							

Figure 4: Accuracy of Feature Pairs (RF Algorithm)

It can be known from Figure 4 that the combination of the perimeter and length of kernel groove attributes of wheat can achieve a better classification effect, and the classification accuracy reaches 96.67%. When using the compactness and asymmetry coefficients for classification, Its classification effect is poor, only 72.38%.

3.2 Naive Bayes classification

Naive Bayes classification is a series of simple probabilistic classifiers based on the application of Bayes' theorem under the assumption of strong (naive) independence among features. This classifier model assigns class labels represented by eigenvalues to problem instances, and the class labels are taken from a finite set. It is not a single algorithm for training this classifier, but a series of algorithms based on the same principle: all Naive Bayes classifiers assume that each feature of

the sample is not related to other features [6].

Assuming that the attributes of any category all follow a Gaussian distribution and the features are independent of each other, pair up the seven attributes. To explore the classification effect, Figures 5 and 6 respectively show the results of combining area with perimeter and length with groove.

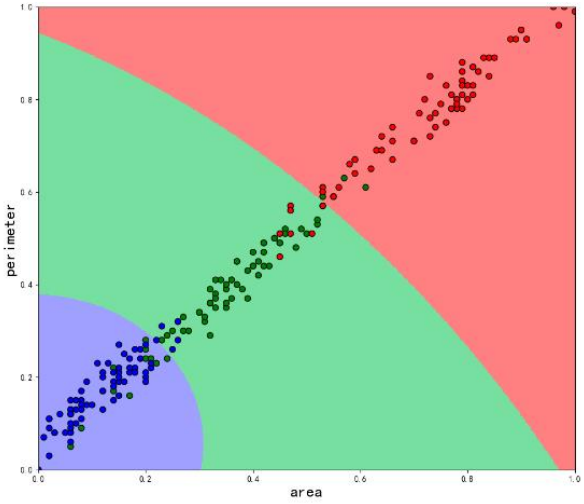

Figure 5: Area + Perimeter

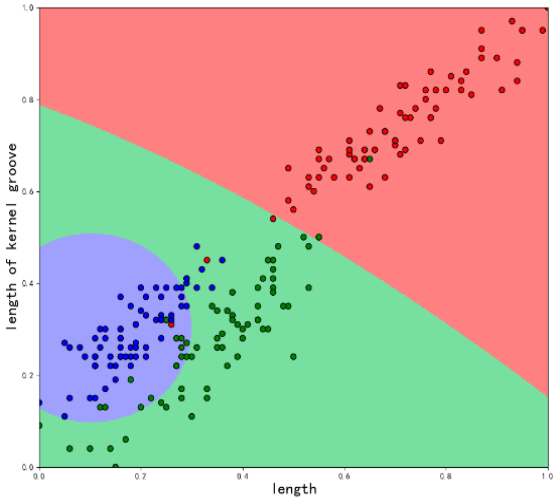

Figure 6: Kernel Length + Groove Length

Among them, Figure 5 shows the combination of the two attributes of area and perimeter of wheat, and its accuracy rate reaches 86.67%. Figure 6 shows the combination of the two attributes of length and length of kernel groove, and its accuracy rate is 86.19%. Next, combine the seven attributes in pairs to obtain the prediction results as shown in Figure 7:

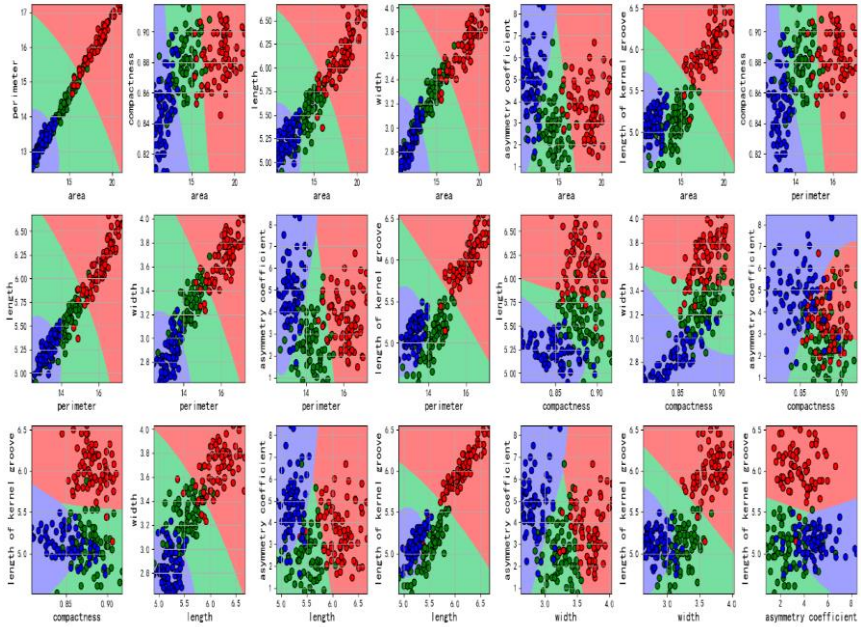

Figure 7 Impact of Feature Pairs on Classification (Naive Bayes Algorithm)

The accuracy rate of its classification results is shown in Figure 8:

According to the relevant data, it can be known that the combination of the perimeter and the length of kernel groove attributes still has a relatively high accuracy, reaching 93.81%. When classifying using the two attributes of compactness and asymmetry coefficient, Its classification

effect is still the worst, with only 70.95%.

Naive Bayes							
Attributes	area	perimeter	compactness	length	width	asymmetry coefficient	length of kernel groove
area		86.67%	86.67%	86.67%	85.71%	91.90%	91.43%
perimeter			86.19%	87.14%	86.19%	90.48%	93.81%
compactness				87.14%	81.90%	70.95%	87.14%
length					84.76%	89.05%	86.19%
width						89.05%	90.00%
asymmetry coefficient							86.19%
length of kernel groove							

Figure 8: Accuracy of Feature Pairs (Naive Bayes Algorithm)

3.3 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a very popular supervised learning algorithm, which can be applied to both linear and nonlinear data. It utilizes a nonlinear transformation to map the original training data onto a high-dimensional space. In the new high-dimensional space, it searches for the linear optimal classification hyperplane, that is, it searches for the separated decision boundaries between one or two different types. Data is mapped to a sufficiently high dimension through nonlinear mapping, and data from two different classes can always be separated by a hyperplane. SVM uses support vectors (basic training tuples) and edges (defined by support vectors) to discover hyperplanes.

Among them, the classification Margin between the two samples is $r = \frac{2}{\|\omega\|}$,The purpose of the support vector machine is to maximize r, which is equivalent to minimizing $\frac{\|w\|}{2}$ or $\frac{\|\omega\|^2}{2}$.

Based on the above theory, we combine the seven attributes of the wheat seed dataset in pairs. The results as shown in Figure 9 were explored:

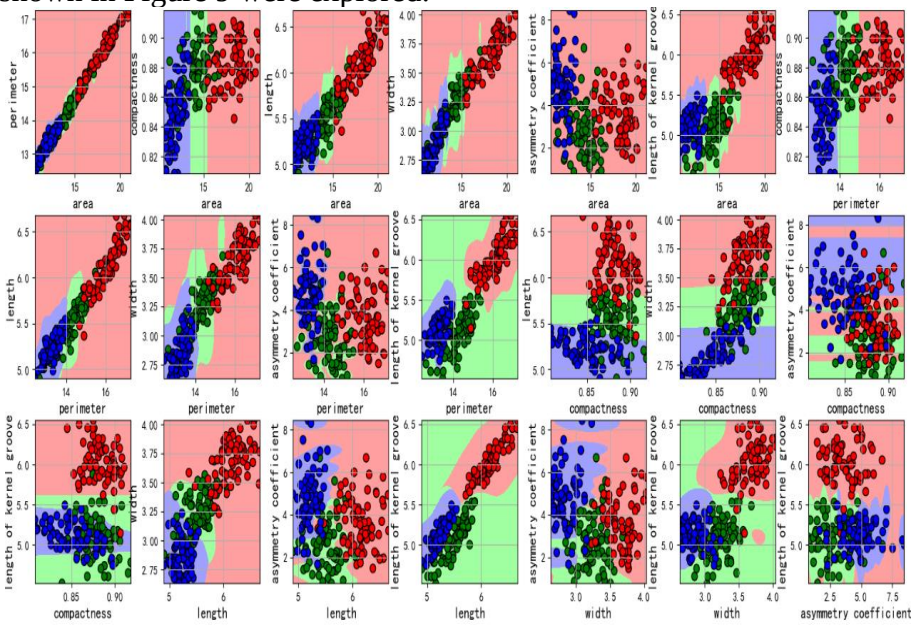

Figure 9: Impact of Feature Pairs on Classification (SVM Algorithm)

The accuracy rate of its classification results is shown in Figure 10:

It can be known from the experimental result data that the area and width of wheat, when combined with the length of kernel groove attribute respectively, have a relatively high accuracy, reaching 97.61%. When classifying using the two attributes of compactness and asymmetry coefficient, its classification effect is still the worst, with only 60.71%.

Support Vector Machine (SVM)							
Attributes	area	perimeter	compactness	length	width	asymmetry coefficient	length of kernel groove
area		90.47%	92.85%	91.66%	86.90%	91.66%	97.61%
perimeter			92.85%	90.47%	90.47%	92.85%	96.42%
compactness				88.09%	90.47%	60.71%	85.71%
length					89.28%	91.66%	95.23%
width						91.66%	97.61%
asymmetry coefficient							88.09%
length of kernel groove							

Figure 10: Accuracy of Feature Pairs (SVM Algorithm)

4. Results Analysis

In this paper, the random forest algorithm, Naive Bayes algorithm and support vector machine algorithm are applied to the wheat seed classification model. According to the output results of each model, they are compared and analyzed, and the following conclusions are drawn:

(1)In terms of classification accuracy, the classification models constructed based on the random forest and support vector machine algorithms are significantly superior to those constructed based on the Naive Bayes algorithm. That is, the classification accuracy obtained by the classification models constructed based on the random forest and support vector machine algorithms is higher than that of the models constructed based on the Naive Bayes algorithm.

(2)Among the first two algorithms, the combination of the two features of wheat perimeter and kernel groove length yields the highest classification accuracy of wheat seeds, which can reach up to 96.67%. In the support vector machine algorithm, the combination of wheat area and width with kernel groove length respectively yields the highest accuracy. Therefore, the feature of wheat kernel groove length can be used as the most important basis for wheat classification. However, the classification effect obtained by combining the compactness of wheat seeds with the asymmetry coefficient was the worst, which was related to their weak correlation.

5. Conclusion

Based on machine learning algorithms and data analysis visualization technology, this study conducted a systematic research on the classification problem of wheat seeds and mainly reached the following conclusions:

(1)In terms of the comparison of algorithm performance, by comparing the classification effects of three algorithms, namely Random Forest, Naive Bayes and Support Vector Machine, it was found that Random Forest and Support vector Machine performed outstandingly in the task of wheat seed classification, and their classification accuracy was significantly higher than that of the Naive Bayes algorithm. Among them, the classification accuracy of the support vector machine under the optimal feature combination reached 97.61%, demonstrating a powerful classification ability.

(2)Regarding the combination of wheat features, the experimental results show that the combination of the length of the nuclear groove with other features (such as perimeter, area, and width) significantly improves the classification effect. Especially, the combination of perimeter and the length of the nuclear groove achieved an accuracy rate of 96.67% in the random forest algorithm, indicating that the length of the nuclear groove is a key feature for wheat seed classification. This discovery provides an important reference for feature selection in practical applications.

(3)In terms of future practical application significance, this study verified the practical value of machine learning algorithms in wheat seed classification, providing a technical solution for automated classification in the agricultural field. By optimizing the combination of features and the

selection of algorithms, the classification efficiency can be significantly improved, the labor cost can be reduced, and it has a positive promoting effect on food security and agricultural modernization.

Acknowledgement

This work was supported by Teaching Reform Project from Haojing College of Shaanxi University of Science & Technology, Teaching Research and Practice of "Data Analysis and Visualization" Course Based on OBE-CDIO Concept in the Context of Interdisciplinary Integration (2024JG050).

This work was supported by Comprehensive Experimental Project from Haojing College of Shaanxi University of Science & Technology, Data Analysis and Visualization of the CDIO Model Based on the Background of Disciplinary Integration (JPSY24005).

This work was supported by Key Research Project on Higher Education Teaching Reform at the University Level in 2024 from Haojing College of Shaanxi University of Science & Technology, Research on AI-Empowered Internship and Practice - Exploration and Implementation of Internship Models to Enhance Students' Professional Competitiveness (2024JG007).

References

[1] Zhao J, Li Q. (2025). Mitigating Distribution Shift in Machine Learning–Augmented Hybrid Simulation [J]. SIAM Journal on Scientific Computing, 47 (2): 475-500.

[2] Sharma, R. (2016). The Role of Big Data Analytics in Business Decision-Making. International Journal of Business Analytics, 14(2), 130-145.

[3] Vakili S, Mousavi M S. (2025). Investigation of the effect of climatic parameters in machine learning algorithms for streamflow predicting in Hamoon Helmand Catchment, Iran [J]. Arabian Journal of Geosciences, 18(5): 106-108.

[4] Eckelt K, Gadhave K, Lex A, et al. (2024). Loops: Leveraging Provenance and Visualization to Support Exploratory Data Analysis in Notebooks. [J]. IEEE transactions on visualization and computer graphics, 4(2), 30-45.

[5] Chen Y. et al. Investigation of the random forest framework for classification of hyperspectral data [J]. IEEE Transactions on Geoscience and Remote Sensing, 2005, 43 (3): 492-501.

[6] Fadlil A, Riadi I, Putra P D J I. Comparison of Machine Learning Performance Using Naive Bayes and Random Forest Methods to Classify Batik Fabric Patterns [J]. Revue d'Intelligence Artificielle, 2023, 37(2): 56-67.