Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 033-040
https://doi.org//10.47514/kjcs/2025.2.1.004

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

An Enhanced Rice Grain Classification Using a Novel Feature Engineering and Random Forest Modelling

Johnson Oluwaseyi Fashanu[1], Usman Ismail Abdulmalik[2] and Mustapha Kassim[3]
[1]Department of Computer Science, Federal Polytechnic, Kaura Namoda, Zamfara State, Nigeria
[2]Department of Computer Science, Federal Polytechnic, Kaura Namoda, Zamfara State, Nigeria
[3]Department of Computer Science, Federal Polytechnic, Mubi, Adamawa State, Nigeria
Johnson.cs@fedponam.edu.ng[1], usmaniamailbox@yahoo.com[2], mustaphakassim@rocketmail.com[3]

Abstract

This research presents a novel approach for rice grain classification, distinguishing between the Cammeo and Osmanick varieties using a dataset obtained from the UCI Machine Learning Repository. The dataset includes morphological features such as area, perimeter, major axis length, minor axis length, and eccentricity. To enhance predictive performance, a Fibonacci-derived linear and polynomial feature engineering technique was introduced, effectively increasing the dataset's feature space. The engineered features were combined and dimensionality reduction was applied using Principal Component Analysis (PCA) with 56 components. Feature selection was conducted using a Random Forest classifier with 500 estimators and a maximum depth of 15. Subsequently, a Random Forest model was employed for classification. The model achieved an accuracy of 93.70% and a mean cross-validation score of 91.34% (5-fold). Class-wise performance metrics for Cammeo included a precision of 0.92, recall of 0.95, and F1-score of 0.93, with support of 350 samples. For Osmanick, the precision was 0.95; recall 0.93, and F1-score 0.94, with support of 412 samples. The confusion matrix highlighted a strong classification performance, with 331 correct classifications for Cammeo and 383 for Osmanick. The dataset was obtained from https://www.muratkoklu.com/datasets/, it is named Rice Dataset (Commeo and Osmancik). This work demonstrates the efficacy of combining this method in enhancing the classification of rice grain varieties.

Keywords: Rice classification, Random forest, Fibonacci sequence, Feature engineering, Polynomial feature

1. Introduction

There is little separation between mathematical sequence and machine learning, this is as a result of recent rise in researchers' effort at intersecting both. This research presents a novel approach for rice grain classification, distinguishing between the Cammeo and Osmanick varieties using a dataset obtained from the UCI Machine Learning Repository. This research aims to explore the integration of Fibonacci-derived polynomial features into machine learning models to enhance the classification performance for rice grains. The unique mathematical properties of Fibonacci sequences provide an opportunity to capture intricate patterns within the grain data that traditional features might miss. The role of rice as a staple food for more than half of the world's population cannot be overemphasized; its quality assessment plays a significant role in global food security and trade. Among the numerous varieties of rice, Cammeo and Osmanick are particularly valued for their distinct physical and culinary properties. Accurate classification of these rice grains is essential for ensuring quality control, preventing adulteration, and meeting market standards. Traditional methods for rice grain classification rely on manual inspection, which is labor-intensive, time-consuming, and prone to human error. Advances in image processing and machine learning offer promising

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 033-040
https://doi.org//10.47514/kjcs/2025.2.1.004

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

alternatives for automating this process, modern classification approaches leverage features such as grain size, shape, color, and texture to differentiate between varieties with high precision. Techniques such as Convolutional Neural Networks (CNNs) have shown remarkable success in object recognition tasks, making them suitable candidates for rice grain classification applications (Abueleiwa & Abu-Naser, 2024). Despite these advancements, research on the classification of Cammeo and Osmanick rice grains remains limited. Existing studies predominantly focus on common varieties such as Basmati, Jasmine, and Japonica, leaving a gap in automated classification systems tailored to regional or less-researched varieties. This study aims to address this gap by developing a robust classification system for Cammeo and Osmanick rice grains using a combination of image processing and machine learning techniques, Random Forest Model was preferred in this research because of its robustness and its feature selection attribute. The Fibonacci sequence, a series of numbers where each number is the sum of the two preceding ones, has intrigued mathematicians and scientists for centuries due to its occurrence in various natural phenomena. Fibonacci numbers and their related polynomial forms have been studied for their unique properties and applications in diverse fields, including computer science, finance, and biology (Livio, 2003).

2. Related Works

A study by Abueleiwa & Abu-Naser (2024) classified five varieties of rice each having 15000 images making a total of 75000 instances in the dataset. They employed deep learning for classification and assessed the model using various metrics, including precision, accuracy, and F1-score. The system was based on a Convolutional Neural Network (CNN) and achieved an impressive accuracy of 99.96%.

Farahnakian et al. (2024) analyzed the performance of deep learning models, including ResNet, VGG, EfficientNet, and MobileNet, on a dataset of 75,000 rice images classified into five categories. The models were evaluated using accuracy, F1 score, precision, recall, and per-class accuracy. EfficientNet achieved the highest accuracy (99.67%), while MobileNet was the fastest (2556 s). The study concluded that these models are highly scalable and effective in handling large, complex datasets compared to traditional machine learning methods.

Tasci et al. (2023) developed Quantized Neural Network (QNN) models to classify five rice varieties using a dataset of 75,000 images from Turkey. Their study focused on reducing computational costs and power consumption, making AI applications feasible on microcontroller-driven edge devices. By implementing eight QNN models based on MLP and Lenet-5 architectures with different quantization levels, they achieved a 99.87% accuracy with the W3A3 Lenet-5-based QNN while using only 23.1 KB of memory. This approach significantly lowered computational demands, with operations per second reduced by 23 times compared to similar studies.

Hazra et al. (2023) examined four CNN architectures VGG-16, VGG-19, ResNet50, and InceptionV3 for classifying five rice varieties: Basmati, Jasmine, Arborio, Karacadag, and Ipsala. The study utilized a dataset of 75,000 images (15,000 per variety) with a uniform size of 250 × 250 pixels. Model performance was assessed using accuracy, confusion matrix, recall, precision, and F1-score. The results highlighted the effectiveness of deep learning models in accurately classifying rice varieties, demonstrating their potential for automating quality control and classification in the rice industry.

Iqbal et al. (2023) explored lightweight deep learning models for rice variety classification, focusing on reducing computational overhead and expert supervision. While traditional deep learning models achieve high accuracy, they demand excessive resources. To address this, the study proposed three models that maintain similar performance while being 10% more efficient. These models were trained end-to-end,

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

minimizing the need for manual preprocessing and feature engineering. The study successfully classified five rice varieties—Arborio, Basmati, Ipsala, Jasmine, and Karacadag—and demonstrated the potential for deploying these models on edge and mobile devices for autonomous field applications.

Saxena et al. (2022) investigated rice variety classification using machine learning algorithms. The dataset included 75,000 samples across five globally grown rice varieties, with 15,000 samples per class. From 107 features, the top 20 were selected using a Random Forest Classifier. Various machine learning models, including logistic regression, decision tree, SVM, random forest, perceptron, KNN, and Gaussian naïve Bayes, were trained on different training-testing splits (70-30% and 80-20%). Performance was evaluated using accuracy, precision, recall, and F1-score. The Random Forest model achieved the highest accuracy (99.85%), followed by the Decision Tree (99.68%), demonstrating the effectiveness of machine learning in rice classification.

Koklu, Cinar & Taspinar (2021) studied five rice Varieties-Arborio, Basmati, Ipsala, Jasmine, and Karacadag commonly grown in Turkey. Their dataset included 75,000 grain images (15,000 per variety) and a second dataset with 106 extracted features, including 12 morphological, 4 shape, and 90 color features. Classification models were developed using ANN and DNN for the feature dataset and CNN for the image dataset. Performance metrics such as sensitivity, specificity, precision, F1-score, accuracy, false positive rate, and false negative rate were analyzed. The classification accuracies achieved were 99.87% for ANN, 99.95% for DNN, and 100% for CNN, demonstrating the effectiveness of these models for rice variety classification.

Cinar & Koklu (2019) used a controlled image acquisition method to capture rice images, placing grains in a light-isolated box to prevent shadows. They studied two rice species, Cammeo and Osmanick, collecting 3,810 images. Morphological features such as area, perimeter, and eccentricity were extracted for classification. Various machine learning models were evaluated, with accuracy rates as follows: Linear Regression (93.02%), Multilayer Perceptron (92.86%), SVM (92.85%), Decision Tree (92.83%), Random Forest (92.49%), Naïve Bayes (91.71%), and K-NN (88.58%). Using an 80-20% train-test split on the Random Forest model, the method achieved a higher accuracy of 93.70%.

This study addresses the gap in automated classification of less-researched rice varieties, specifically Cammeo and Osmanick, by introducing a novel feature engineering approach using Fibonacci-derived polynomial transformations. While existing research primarily focuses on common varieties like Basmati and Jasmine, this study introduce the integration of mathematical sequence-based features to enhance classification accuracy. Traditional methods rely on morphological and colour features, but they often fail to capture non-linear relationships within the data. By incorporating Fibonacci-based transformations and leveraging Principal Component Analysis (PCA) for dimensionality reduction, the study optimizes feature selection and improves model performance. Furthermore, unlike prior works that predominantly use deep learning models like CNNs, this research employs a Random Forest classifier with enhanced feature representation, achieving a significant improvement in classification accuracy.

2.1 Framework Design

An innovative technique which incorporates the use of deterministic mathematic equations for feature engineering was employed; the approach broadened the number of features from which the machine learning model can select from, Random Forest feature selection was used in this study, this approach leverages from the fact that it certainly captures non-linear relationships among features, hence enhancing the performance of the model. Some of the benefits of this approach are highlighted below:

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 033-040
https://doi.org//10.47514/kjcs/2025.2.1.004

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

2.1.1 Improved Performance

The incorporation of Fibonacci-derived features enhances the performance of the rice classification model. The combination of the original and derived features increased the features whose dimension was reduced before features selection. It achieved a higher performance.

2.1.2 Novel Application

While Fibonacci sequences have been extensively studied in mathematics and applied in various fields such as biology and finance (Livio, 2003), their application in feature engineering for machine learning is relatively unexplored. This study pioneers this novel application, opening new avenues for research and innovation.

3. Methodology

Machine Learning (ML) is the programming of a computer to make prediction using data provided. It is a branch of artificial intelligence (AI) that develops algorithms that enable computers to learn from data and make predictions or decisions without being explicitly programmed (Nelson, 2020). The model may be predictive to make predictions in the future, or descriptive to gain knowledge from data, or both. This research aims to develop a machine learning framework that integrates Fibonacci-derived polynomial features for classification modeling of rice varieties (Cammeo and Osmanick). The methodology includes data collection, data preprocessing, feature engineering, model application, evaluation, and validation using cross-validation techniques. The study follows a structured approach to ensure rigorous analysis and reliable results.Figure1shows the framework of the model.

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 033-040
https://doi.org//10.47514/kjcs/2025.2.1.004

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

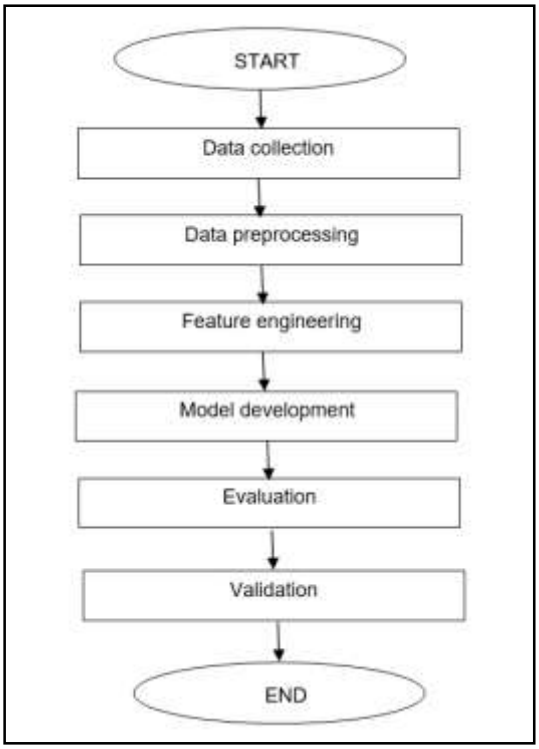

Figure 1: Framework of the model

4. Results and discussion

The result of this research work will be presented in this section, the implementation of feature engineering techniques using polynomial features yielded significant insights and classification capabilities for the rice varieties. The combination of the features increased the achievement in terms of accuracy; this shows that non-linear relationships among these features play a significant role. Dimensionality reduction was applied using Principal Component Analysis (PCA) with 56 components. PCA is used to reduce the dimensionality of datasets by transforming features into a set of linearly uncorrelated components. PCA was used because of the many features resulting from engineering. The table 1 shows how much influence the engineered features had on the model.

Table 1: Five Most Influential Features for Five Principal Components

	1st Influential Feature	**2nd Influential Feature**	**3rd Influential Feature**	**4th Influential Feature**	**5th Influential Feature**
PC1	ConvexArea poly7	Area poly7	ConvexArea poly6	Area poly6	ConvexArea poly5
PC2	Area poly7	ConvexArea poly7	ConvexArea poly6	Area poly6	ConvexArea poly5
PC3	ConvexArea poly6	Area poly6	ConvexArea poly5	Area poly5	ConvexArea poly7

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 033-040
https://doi.org//10.47514/kjcs/2025.2.1.004

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

PC4	Area poly6	ConvexArea poly6	Perimeter linear feature	MinorAxisLength linear feature	Extent
PC5	Perimeter linear feature	Extent	MajorAxisLength linear feature	Area linear feature	MinorAxisLength linear feature

Feature selection was conducted using a Random Forest classifier with 500 estimators and a maximum depth of 15. Subsequently, a Random Forest model was employed for classification. The model achieved an accuracy of 93.70% and a mean cross-validation score of 91.34% (5-fold). Figure 2 shows the fivefold cross validation.

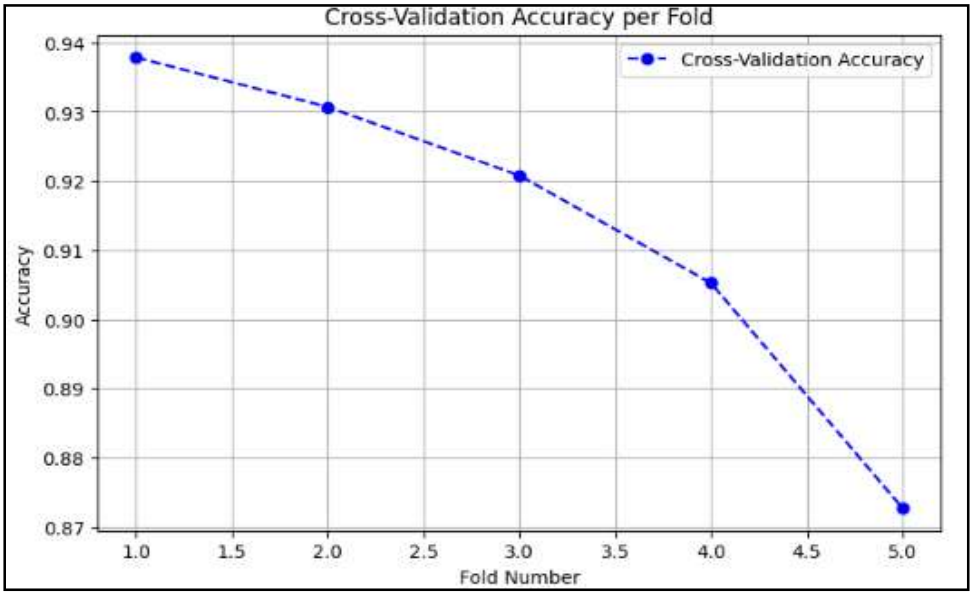

Fig 2: Graph of Fivefold Cross Validation

A class-wise performance metrics for Cammeo and Osmanick were carried out, Cammeo was classified as 0 and Osmanick as 1. Precision, recall, and F1-score are evaluated with support of 762 samples as illustrated in Table 2.

Table 2: Performance Metrics of Model

Rice Varieties	**Precision**	**Recall**	**f1-score**	**support**
0	0.92	0.95	0.93	350
1	0.95	0.93	0.94	412
Accuracy			0.94	762
Macro avg	0.94	0.94	0.94	762
Weighted avg	0.94	0.94	0.94	762

The confusion matrix highlighted a strong classification performance, with 331 correct classifications for Cammeo and 383 for Osmanick. Figure 3 displays the confusion matrix.

Kasu Journal of Computer Science Vol.2 No.1 [March, 2025], pp. 033-040
https://doi.org//10.47514/kjcs/2025.2.1.004

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

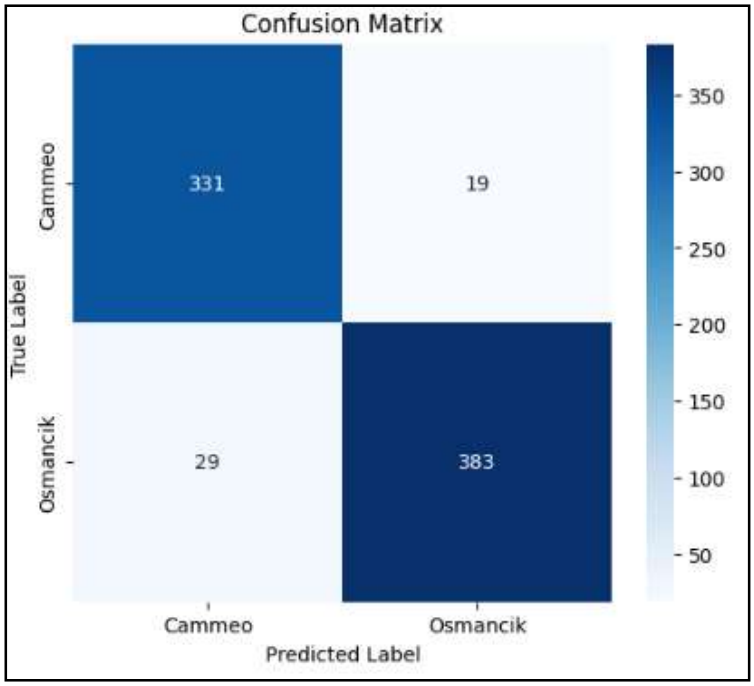

Figure 3: Measurement of Confusion Matrix

The results of this study indicate that the proposed Fibonacci-derived feature engineering technique, combined with Principal Component Analysis (PCA) and Random Forest modeling, significantly enhances rice grain classification accuracy. The classification model, trained on a feature-enhanced dataset, achieved an impressive accuracy of 93.70%, with a mean cross-validation score of 91.34% across five folds, demonstrating the model's strong generalization capability. The introduction of Fibonacci-derived linear and polynomial transformations effectively expanded the feature space, allowing the model to capture subtle morphological differences between rice varieties. Traditional morphological feature-based classification may achieve decent accuracy, but the incorporation of Fibonacci-derived transformations and PCA provides a systematic enhancement by generating richer feature representations. The preprocessing of the dataset which expanded features as well as well the application PCA to reduce dimension of the dataset improved the performance of the model.

5. Conclusion

This study focus on the impact of dimensional reduction and the influence of PCA on the performance of the model, however correlation matrix of the features were not evaluated, and further research can be carried out other than using PCA such as Variance Inflation Factor (VIF) to select features. Polynomial Features tend to have multicollinearity which details high correlation amongst features which may affect the performance of model with a little change in the dataset; VIF reduces multicollinearity by dropping feature which could lead to better performance and a stable the model.

Online ISSN: 1597-2216
Print ISSN: 1597-2178
www.kjcs.edu.ng

References

Abueleiwa, M. and Abu-Naser, S. S. (2024). Classification of rice using deep learning. *International Journal of Academic Information Systems Research (IJAISR)*, 8(4), 26-36. Retrieved from www.ijeais.org/ijaisr

Cinar, I. and Koklu, M. (2019). Classification of rice varieties using artificial intelligence methods. *International Journal of Intelligent Systems and Applications in Engineering*, 7(3) 188-194. https://doi.org/10.18201/ijisae.2019355381.

Farahnakian, F., Sheikh, J., Farahnakian, F. and Heikkonen, J. (2024). A comparative study of state of the art deep learning architectures for rice grain classification. *Journal of Agriculture and Food Research*, 15 (2024). https://doi.org/10.1016/j.jafr.2023.100890

Hazra, S., Karforma, S. Bandyopadhyay, A., Banerjee, A., Chakraborty, D. (2023). Deep learning-based classification of rice varieties: A comprehensive study. *Journal of Survey in Fisheries Sciences*, 10(3) 411-419. Retrieved from https://www.researchgate.net/publication/385142789

Iqbal, M. J., Aasem, M., Ahmad, I., Alassafi, M. O., Bakhsh, S. T., Noreen, A., and Alhomoud, A. (2023). On application of lightweight models for rice variety classification and their potential in edge computing. *Foods, MPDI*, 2023, 12(21), 1-16. https://doi.org/10.3390/foods12213993

Koklu, M., Cinar, I., Taspinar, Y. S. (2021). Classification of rice varieties with deep learning methods. *Computers and Electronics in Agriculture, Elsevier*, 187 (2021). https://doi.org/10.1016/j.compag.2021.106285

Livio, M. (2003). The golden ratio: The story of Phi. *The World's Most Astonishing Number. New York, NY*: Broadway Books.

Nelson, D. (2020). What is Ensemble Learning? Retrieved November 20, 2023, from https://www.unite.ai/what-is-ensemble-learning/

Saxena, P., Priya, K., Goel, S., Aggarwal, P. K., Sinha, A. and Jain, P. (2022). Rice varieties classification using machine learning algorithms. *Journal of Pharmaceutical Negative Results*. 13(7), 3762-3772. DOI: 10.47750/pnr.2022.13.S07.479

Tasci, M., Istanbullu, A., Kosunalp, S., Iliev, T., Stoyanov, I., Beloev, I. (2023). An efficient classification of rice variety with quantized neural networks. *Electronics, MPDI*, 12(2285), 1-16. https://doi.org/10.3390/electronics12102285