

Hybrid Machine Learning Merging KNN and Naive Bayes Weighted Polling for Dry Beans Multiclass Classification

Farid Morsidi*

Universiti Pendidikan Sultan Idris, Faculty of Computing and Meta-Technology, Computing Department, 35900 Tanjong Malim, Perak, Malaysia.

Multiclass classification is essential in machine learning as it allows systems to differentiate between three or more categories using intricate, multidimensional data. This paper examines the theoretical basis and real-world applications of multiclass classification, highlighting feature selection methods and the efficacy of classifiers. Essential principles, including significant inter-class variance, minimal inter-feature correlation, robust target relevance, and feature consistency, are examined for their contributions to discovering optimal predictors across various datasets. The experimentation is implemented with a dry bean dataset, utilising classifiers such as K-Nearest Neighbors (KNN) and Naive Bayes to distinguish bean types based on their morphological characteristics. KNN excels in basic pattern recognition, and Naive Bayes implements classification probabilistically, which is appropriate, widespread and accurate in high-dimensional environments. The study also stresses the importance of performance metrics such as accuracy, and dimensionality reduction as useful analytic tools. The method of feature selection and classifier combined achieved the accuracy measures from the lowest value of 85.67% to the highest value of 88.58%. Although these results look good for the automated classification of beans, performing thorough evaluation with balanced metrics for all classes would increase the trust in the robustness of the technique, especially for the minority bean types.

Keywords: instance-based learning; probability classification; machine learning; KNN; Naïve Bayes

I. INTRODUCTION

Multiclass classification is a core problem within machine learning, in which the learning algorithm learns to classify one input into one of many class instances (Sarki *et al.*, 2022). The distinction is that there is a yes/no or this/that decision in binary classification; multiclass classification denotes determining which of the distinctive category makes up the entire domain. The notion of multiclass classification is to describe aspects of reality, the complexity of reference frames, and when to make calls between choices that might include distinctions beyond a yes/no distinction (Shamrat *et al.*, 2023; Adi Pratama *et al.*, 2024). The relevance of multiclass classification to machine learning is when specifically addressing the multiclass classification embedded in the

definition of classifying instances belonging to more than the two classes being considered. The component of a multiclass classification emerges sort of differently than that of a binary classification, as their construction level is at a real-world level, and they account for complexity and making distinctions among multiple classes, not just as a binary classification (Wang, Muthukumar and Thrampoulidis, 2023). When classification involves noticing minor differences between the feature traits which distinguish between each of the classes being measured, the models should be trained across multiple categories to account for multiclass classifications. The models become more flexible to new data, and generalise much better as a concept, especially as new classes are added.

*Corresponding author's e-mail: farid.mors90@gmail.com

Feature selection, a well-known metric classification measure naturally originating from machine learning as a metric for improved class segregation, relied excessively upon multiclass classification to evaluate which features had greater discriminative power across all target classes, rather than two separate independent classes (Rosly *et al.*, 2024). This allowed researchers to highlight features that had far greater variation across classes without tightly adhering to pairwise differences that lead to overfitting. Furthermore, multiclass classification provides an iterative feedback loop whereby the classification assists improvements in feature engineering for example, investigating confusion matrices or determining where specific classes had low recall (Shamrat *et al.*, 2023). Misclassifications and identifying missing or weak features are often routine considerations in multiclass classification; therefore, it should encompass a deeper evaluation of more complex feature transformations or the addition of extra variables to enhance the evaluation metric to gain an improved understanding on a classification systems' capabilities. Evaluation strategies such as One-vs-Rest (OvR), One-vs-One (OvO) or native multiclass models such as decision trees, random forests and neural networks are used for the implementation of multiclass classification (Moral, Nowaczyk and Pashami, 2022; Wang, Muthukumar and Thrampoulidis, 2023; Morsidi, 2025). The training data allows the algorithms to learn a mapping occur between the input features and class labels, which is the result of learning the patterns that occur. In preparation for predicting unseen data, the evaluation on predicting class for the probability of each class and assigns the input to that class with the maximum confidence. Some of the most utilised evaluation measures include accuracy, precision, recall, F1-score, and ROC-AUC per class, to ensure good performance across classification (Javed Mehedi Shamrat *et al.*, 2022; Chabalala, Adam and Ali, 2023). Multiclass classification presents a detailed improvement in predicting outcomes, while also fostering a more informed and comprehensive selection of the feature set.

By employing multiclass classification, it is possible to predict positive outcomes in datasets with various attributes, such as those in agriculture, environmental science, or biology, by analysing them using a complex set of characteristics. Various morphological, chemical, and physiological characteristics can be observed in every sample within agricultural environments, for instance like a bean, fruit, or grain. Various research efforts utilising machine learning have been undertaken for agricultural objectives to

deepen knowledge of the ecosystem and to support quantitative analysis for better agricultural planning. Several common factors considered for categorising plant classes involve size, shape, colour, moisture levels, protein content, and disease resistance (Opara *et al.*, 2024). Using multiclass classification models can aid in identifying or classifying crop varieties, as well as in pinpointing the type most suitable for achieving specific desired outcomes. These include health benefits, economic value, and optimal growing conditions. Agriculture and food technology utilise crop classification, plant disease classifications, or seed quality to enable automated sorting, targeted treatment, precision farming, and increased productivity (Alsmadi and Almarashdeh, 2022; Chabalala, Adam and Ali, 2023). Health standards can be established by examining factors like high protein levels, low fat content, fiber consumption, antioxidant intake, or low levels of harmful substances. These criteria may be used to determine individual health. The nutritional profiles of bean types could be associated with a multiclass classifier that has been trained on these traits. Based on the correlative analysis for corresponding desired traits, it is possible to consistently forecast specific plant species within their natural range possessing the most beneficial health characteristics relative to their high fiber content and low-calorie consumption. For instance, using imaging data to identify disease-resistant traits like shape deformities, leaf color, or pest marks, similar cases could be used to determine the most diseases beans category.

This study attempts to further invigorate the benefits of multiclass classification for agricultural purposes especially in terms of predicting and identifying the best plant specimen respective from their habitat domain, where hybridization of popular parametric modelling approaches for learning algorithms is merged with the intention of improvising their individual implementation within their functional conjectures. Boosting the probabilistic classification properties of instance-based parametric algorithm can be seen fruitful for niche purposes of ideal specimen classifications and segregation of best individual plant classes among their distinctive categories (Chabalala, Adam and Ali, 2023; Morsidi, 2025). For this purpose, dry bean dataset is being referenced as the experimentation tool and proof-of-concept implementations of the proposed classification strategy, where the performance metric would be determined by the relative accuracy for each subsequent classification phases.

This paper is structured as follows: Section 2 discusses the accomplished work in instance-based parametric modelling of multiclass classification strategy within real-world implementations, where Section 3 provides a critical discussion on the feature extractions and implementation strategy conducted during the experimentation on identification of best dry bean samples, Section 4 relays the output and further analysis on the conducted experimentation methods, and Section 5 is the conclusion representing the entire paper.

II. RELATED WORKS

Multiclass classification is an important, supervised machine learning method that involves an algorithm learning from data that is then placed into three or more exclusive classes. Multiclass classification is particularly effective in problems with categorical diversity and high dimensional features (Javed Mehedi Shamrat *et al.*, 2022; Sarki *et al.*, 2022). Image recognition, natural language processing, speech recognition, and bioinformatics, among others, are typical areas where multiclass classification is leveraged. The list of algorithms that can be used for multiclass problems includes multinomial logistic regression, decision trees, random forests, support vector machines, in addition to deep learning methods like convolutional neural networks and recurrent neural networks (Bacevicius and Paulauskaite-Taraseviciene, 2023; Wang, Muthukumar and Thrampoulidis, 2023; Demir and Yaman, 2024). As computational power and data volumes continue to increase, multiclass classification should help improve predictive modelling, adaptive learning, and decision support problem solving across applications.

Multiclass classification in medical imaging and diagnostics differentiate disease types like skin lesions and tumours. Diagnostic aids can automise functions beyond most manual labour practices improving the accuracy of acute detection of significant engagement and detection of Diabetic Eye Disease (DED) to prevent vision loss when visibly affect patients (Sarki *et al.*, 2022). Retinal fundus image screening is also manual in methodology which promotes long screening turnaround time for diseases relying on prompt examination for successful treatment (Sarki *et al.*, 2022). The relevance of multiclass classification in medical purposes is subsequently even more indicative of practice and implication in eye care, given how retinal diseases achieve

utmost representation of the sample manipulation through an artfully depicted manner in addition to improvements on outreach by means of public engagement. In contrast to conventional metrics during the service of reaching blatantly healthy versus diseased images as discoverable with deep learning performance in single layer models such as image classification is elementary to feature dependent technicalities that actualise after DED selection which both confidentiality and medical practice do well to maintain. For instance related with this example, a study was developed to build an automated classification framework for multiclass DED (Sarki *et al.*, 2022). A convoluted neural network model on retinal fundus images was trained from a publicly accessible dataset returning a high max accuracy thus far of 81.33% then without deviation produced 100% sensitivity and specificity.

Manufacturing processes depend on multiclass models for defect classification and quality control. By classifying a defect in a product into multiple categories, an automated system can instruct robotic arms or inspection systems for the correct response. The literature around "benign overfitting" tends to focus on regression or binary classification problems, neglecting multiclass classification where the power of modern machine learning lies (Wang, Muthukumar and Thrampoulidis, 2023). The current study moves to study benign overfitting in a multiclass linear classification context and studies three algorithms: (i) empirical risk minimisation (ERM) with multi-class cross-entropy loss, which converge to multi-class SVM solutions, (ii) ERM with least-squares loss which converges to min-norm interpolating (MNI) solutions, and, (iii) one-vs-all SVM classifiers (Wang, Muthukumar and Thrampoulidis, 2023). Findings reveal that under specific conditions, these algorithms interpolate training data, achieving equal accuracy and demonstrating benign overfitting with sufficient overparameterisation. Novel error bounds for MNI classifiers are also derived, indicating good generalisation for SVM solutions beyond typical margin-based bounds (Wang, Muthukumar and Thrampoulidis, 2023). Assessing ordinal classifiers' performance amidst imbalanced data poses significant challenges. The choice of metrics is crucial, with Cohen's kappa commonly used, but other agreement measures may perform better with various ordinal confusion matrices. This research employs weighted

agreement measures as evaluation metrics for ordinal classifiers, assessing their applicability across diverse fields with challenging data compositions (Yilmaz and Demirhan, 2023). The study analyses the metrics' sensitivity in distinguishing between classifiers, utilising five kappa-like measures with six weighting schemes, where comparisons are done on reliability to mainstream metrics like F1 and Matthews correlation coefficient, evaluating 37 metrics in numerical studies of synthetic and real datasets. The analysis identifies promising metrics and recommends the weighted Scott's pi-measure for its sensitivity to performance differences and reliability. Various machine learning algorithms have been employed for network intrusion classification, addressing both binary and multi-class challenges (Bacevicius and Paulauskaite-Taraseviciene, 2023). Many studies have examined unbalanced datasets such as CIC-IDS2017, mostly by either merging classes or subsampling the dominant class. However, doing this creates a trade-off and more classes do not always affect accuracy negatively. Since a cybersecurity expert needs to pinpoint the attacks and how they occurred, the performance of multi-class classification in the context of imbalanced data from CIC-IDS2017 and CSE-CIC-IDS2018 was examined in a relative study, with models including Logistic Regression and Decision Trees. Using a CART based decision tree, the computation on an impressive average macro F1-score of 0.96878 with 28-class classification was performed in this process.

Multiclass classification can effectively relate to the automation of complex tasks that typically require human review. The application of multiclass classification as part of an automated workflow that contains sensory data, image processing, or natural language increases the complexity of the system, potentially allowing for more robust automation by utilising augmented and intelligent systems to perform inherently complex tasks with reduced human supervision (Sebák and Kacsuk, 2021). A neural network is used to evaluate the diagnosis of various lung diseases such as emphysema or pneumonia from the ChestX-ray14 dataset (Shamrat *et al.*, 2023). The optimised MobileLungNetV2's analysis is completed after preprocessing (CLAHE and Gaussian filter). Compared were multiple transfer learning models to achieve 91.6% accuracy with MobileNetV2. The

optimised model has accuracy of 96.97%, and has very good precision, recall, and specificity scores, which also demonstrates reasonably good efficiency for lesion classification. Another instance of multiclass classification study with probability classification is done on widespread disease generalised by population majority, such as thyroid disease that is considered as a prevalent health issue in India affecting around 42 million individuals. Key thyroid hormones like TSH, T3, FT4, T4, and FTI are crucial for diagnosing conditions such as hypothyroidism and hyperthyroidism, where the feat of analysing these parameters manually in large datasets is challenging. Another relative study utilises various machine learning techniques, including decision trees, random forests, naive Bayes, multiclass classifiers, and an Artificial Neural Network (ANN) for the prediction of hypothyroidism (Guleria *et al.*, 2022). It is very difficult to accurately analyse diagnostic characteristic parameters manually in large data. A multiclass classification implementation study has employed several machine learning processes, such as decision trees, random forests, naive Bayes, multiclass classifiers, and an Artificial Neural Network (ANN) to produce hypotheses related to hypothyroidism. The method resulted in very high accuracy rates of 99.5758% and 99.3107% respectively.

In addition to agriculture, similar methods are used in other fields. In horticulture, multiclass models are employed to help select the right variety of plants for soils or climates (Opara *et al.*, 2024). The major benefit of using multiclass classification is the ability to model and compare many classes at the same time. Typically, researchers want to know which of the competing groups is behaving in a manner of potential interest in comparison to all the other options. The use of raw multidimensional data can thus be translated into desired actionable knowledge (Guleria *et al.*, 2022). In a sustainable agricultural system characterised increasingly by precision farming and data-informed breeding programs; multiclass classification will be ideal for aiding in selection in order to increase yield and nutritional enhancements. In South Africa, fruit tree crops provide food, income opportunities, and contribute significantly to GDP (Chabalala, Adam and Ali, 2023). As such, advances in phenotyping are critical to building precision agriculture in the horticulture industry. Traditional methods of classifying crops have

proven inefficient for heterogeneous systems of fruit-tree species; however, remote sensing shows great potential for advancement. It is vital to note that intercropping and spatial dispersion in the field can create unbalanced representation of composition, impacting classification accuracy even through remote sensing. To counter this issue, a study was developed aiming to assess the impact of balanced and unbalanced multi-class distributions on detection accuracy using a suite of machine learning classifiers with Sentinel-2 (S2) imagery (Chabalala, Adam and Ali, 2023). Of the machine learning algorithms, the SVM algorithm produced the highest accuracy (71%) demonstrating the possibility of effective fruit tree species mapping and management from S2 data.

Multiclass classification in animal breeding can help determine which breeds of livestock are best suited to produce meat, milk, or even if they will be resilient to climate change, in the study of various physiological traits (Suryawanshi and Dutta, 2022). A relative examination of multiclass classification has studied machine learning (ML) applications in horticulture as it relates to predicting losses and waste development with fresh crops. As there is ample opportunity for health with fresh fruit, the application of machine learning to optimise horticulture and agricultural practices is becoming a point of research focus. Whenever production systems can improve, it is an important and valuable opportunity that affects sustainable food and nutrition security. A research review assesses the role of ML in preharvest and postharvest areas, highlighting its ability to minimise postharvest losses by measuring their magnitude, which is essential for effective management and policy development (Opara *et al.*, 2024). This study investigates how machine learning can be applied in various stages of agriculture, from preharvest activities to postharvest management and even waste quantification. The results show that several ML algorithms perform well in classification and predictive tasks. This suggests a good probability of future ventures into combining different models to boost the effectiveness of multiclass classification in agriculture. Future identified research from this study proposed possible exploration on how ML can help minimise waste in these processes.

III. RESEARCH FRAMEWORK

The section covers the use of the proposed classification prediction methods on a selected dataset consisting of seven registered types of dry beans that possess similar traits to achieve uniform seed classification. The discussion emphasises both the characteristics and methods of each classification algorithm, as well as its relative value and importance in relation to classification analysis. This research employs weighted polling to combine the KNN algorithm and Bayes classification, aiming to examine the degree of prediction accuracy resulting from the application of weighted polling to the current frameworks of both algorithms. The research scope explores the benefits of incorporating weighted polling techniques as a selection criterion to identify the most precise classification features from the training dataset in relation to the known output for segmenting and extracting features of each dry bean trait and characteristic in the testing domain.

The present research intentionally uses classic machine learning techniques—K-Nearest Neighbors (KNN) and Naïve Bayes—for a number of strategic purposes. Primarily, in agricultural applications, where breeders and quality controllers need to understand the reasons behind the classification to directly influence the respective programs, interpretability is very important. In contrast to the black-box models like deep neural networks or complex ensemble methods, KNN analogously gives reasoning based on instances that links directly to the similar training examples, while Naïve Bayes lays out probabilistic reasons that rely on feature likelihoods. Secondly, computational efficiency is the decisive factor for the deployment of these methods in the agricultural resources-limited settings, where the real-time classification might be going on edge devices or mobile platforms. Both KNN and Naïve Bayes are very low in computational resource requirements as compared to tree-based ensembles or transformer models. Thirdly, these techniques can perform well with small-to-moderate-sized datasets (as in the case of the 13,611 samples imposed in this study) without requiring extensive hyperparameter tuning or large-scale data augmentation. Lastly, the combining of the probabilistic nature of Naïve Bayes with the local decision boundaries of KNN creates complementary advantages that when correctly weighted can cover both global distributional

patterns and local geometric structures in the feature space—a hybrid advantage that is not easily reached by single-method approaches.

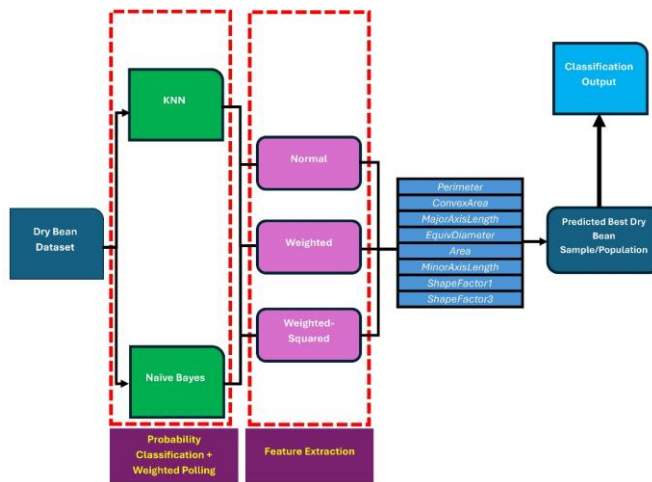


Figure 1. Structure for the executed classification approach for dry beans dataset categorisation utilising KNN and naïve Bayes' weighted voting

The diagram in Figure 1 shows the comprehensive process for the suggested classification framework. First, the dry bean dataset is presented and through feature extraction and selection, the chosen features are extracted. The features are then input to the KNN and Naïve Bayes classifiers which are working at the same time. The KNN part is identifying the local neighbourhoods, and the Naïve Bayes is training in the local subsets. After that, their probability outputs are merged using the weighted polling method, which results in the final classification being made.

A. Dataset Overview

This study used a dataset for dry beans in Koklu's research project (Koklu and Ozkan, 2020), to look at seven different types of dry bean, on shape, typology, and structure based on market conditions. For a uniform seed classification, the study uses three registered types of dry beans with similar characteristics using the dataset that is referenced. Koklu distributed high-resolution camera images with 13,611 grains of 7 different registered dry beans to use for classification. The computer vision system took segmented and pulled features from the beans, which included 16 features that examined the beans across 12 dimensions and three shapes. For example, the Dry Bean dataset is a morphological data set that includes different shape and size features pulled from grains. The dataset consists of seven different types, which

are BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ, SEKER, and SIRA. These represent a variety of different samples from variety of beans. This analysis is at identifying which features are important, evaluates classifiers, and visualise how different types of beans segregate. The rows show a sample of beans, and the columns consist of morphological, and shape features taken from a sample of those beans. The last column is where the class label of a bean type ends.

B. Principled Feature Selection for Multiclass Classification

For multiclass classification, an extensive generalisation can only occur when appropriate features are selected that are representative for every class in the relative domain. To enhance multiclass identification techniques, identification efforts are emphasised on features to extract that distinguish features from one category, as well as features in that group that retain a common format in each group. Advanced methods are typically used for feature selection to distinguish class clusters to make the model more general. This test ensures that features selected for each class are sufficiently distinct such that they won't aid in the classification of categories but serve as model scaffolding around the core class. This inquiry had many objectives to be carried out using the intended process of a multi-class classification.

- i. **Identifying the Best Features in Multiclass Classification:** In multiclass classification tasks, selecting suitable features is essential for creating accurate and interpretable models. Unlike binary classification, which distinguishes two classes, multiclass problems require features that effectively separate multiple categories. This involves evaluating statistical and domain-specific criteria for impactful variable selection (Seb ak and Kacsuk, 2021).
- ii. **High Variance Across Classes:** The idea behind multiclass classification is to keep intra-class variance low, while maximising inter-class variance so that separation into classes can happen clearly (Kuru and Khan, 2018). An agricultural dataset containing different classes of beans was referenced as the main example in this study, highlighting a crucial feature: *Perimeter*. With *Perimeter* values

clustered tightly together, the decision boundaries are sharpened to reduce the chances of making errors when classifying the objects. Intra-class variance gets in the way of making successful classifications. A critical point of this discussion is that successful classification is a function of adequate separation of classes and having minimal noise within each of the classes.

- iii. **Low Correlation with Other Features:** In multiclass classification, highly correlated features can increase model complexity and lead to overfitting and multicollinearity, especially in linear models like logistic regression. For instance, in shape-based datasets referenced in the study that is the dry bean feature datasets, features such as *Area* and *ConvexArea* may provide redundant information, which doesn't enhance classification performance. Reducing feature redundancy improves computational efficiency, interpretability, and model stability, ensuring independent contributions from each feature. This study explores instance-based learning with parametric classifiers, focusing on majority vote polling among the closest neighbours to achieve optimal accuracy rates across multiclass clusters of varying samples.
- iv. **High Correlation with the Target Class:** In multiclass classification, the best predictors correlate strongly with the target variable to differentiate among the class labels. Correlation methods typically do not work for a categorical target, hence researchers applied instance-based learning and looked for discriminative metrics to evaluate relevance of the predictors from the target variable. These methods identified important predictors that contributed to separation of the classes based on both linear and non-linear relationships. The study focused on relevance of predictors combined across classes to prevent bias towards the dominate categories and improve classification strategies through analysis of correlations in the multiclass clusters.
- v. **Consistency Across Samples:** Stability and reliability of features are critical to the performance of models developed with them, as it allows robust

predictions to occur upon introducing new data. The general expectation is that real-world situations support consistent features for classification. To identify feature stability, several evaluation tools were incorporated such as outlier detection, variance thresholds, and distribution analysis. Features classified with a high degree of noise should be transformed or eliminated. A process of systematic statistical procedures to support multiclass classification leads to better overall outcomes by prioritising consistent features and standard deviations between classes. Determining the stability of a distribution within classes supports the determination of a unique feature set which discriminates the performance of the model and its overall generalisation.

The following table illustrates several parameter variables chosen as the best feature selection characteristics representing the entire dry bean dataset according to their relevancy with each other aside from the congruity of the accuracy values associated with each class cluster.

Table 1. Attribute Information for the Referenced Multiclass Dataset (Dry Bean)

Item	Description
<i>Perimeter</i>	Bean circumference is defined as the length of its border.
<i>ConvexArea</i>	Number of pixels in the smallest convex polygon that can contain the area of a bean seed
<i>MajorAxisLength</i>	The distance between the ends of the longest line that can be drawn from a bean
<i>EquivDiameter</i>	The diameter of a circle having the same area as a bean seed area
<i>Area</i>	The area of a bean zone and the number of pixels within its boundaries

<i>MinorAxisLength</i>	The longest line that can be drawn from the bean while standing perpendicular to the main axis
<i>ShapeFactor1</i>	Statistical shape descriptors
<i>ShapeFactor3</i>	Statistical shape descriptors

These features capture shape geometry in diverse ways, making them valuable for distinguishing bean types. This multiclass classification experimentation tested a total of 8 distinctive features constituted by 7 different dry bean classes.

Table 2. Class Types from the Multiclass Dataset (Dry Bean)

Bean Class	Bean Type
0	BARBUNYA
1	BOMBAY
2	CALI
3	DERMASON
4	HOROZ
5	SEKER
6	SIRA

C. Multiclass Classification Ensemble with Weighted Polling combining KNN & Naïve Bayes

1. K-Nearest Neighbour (KNN)

KNN is a supervised machine learning method developed by Evelyn Fix and Joseph Hodges in 1951 (Fondón *et al.*, 2018). KNN is a respected dynamically referenced approach in classification and regression analysis, as it avoids assumptions about data distribution. KNN's evolving characteristics allow it to handle alphanumeric data types, making it suitable for various dataset categories (Vidyarthi *et al.*, 2022; Dehkordi, Manochehri and Aghazarian, 2023). Its lower sensitivity to outliers makes it a viable option. KNN starts by searching for the unique number of neighbouring clusters in relation to a given data point, using distance metric techniques defined by the classification type. Pre-processing techniques like weighting and dimensionality reduction enhance accuracy and predict the viability of the local data structure.

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2} \quad (1)$$

KNN is a machine learning algorithm that assigns an unclassified object to a binding class about its k nearest neighbours within the training set (Vidyarthi *et al.*, 2022; Taylor, 2023; Morsidi, 2025). This approach ensures preferable asymptotic performance with an infinite optimal Bayesian error rate. The weighted KNN rule was proposed to address trade-offs in contemporary nearest neighbour classification. The sensitivity of k neighbourhood size is directly influenced by the selection of optimal k values, with a minute value producing below-average local estimates and vice versa (Dehkordi, Manochehri and Aghazarian, 2023; Salauddin Khan *et al.*, 2023). Feature extraction is used to complement KNN's deficiencies in improving output performance, even when data points are not gleaned, such as mislabelled, vague, or outlier-filled data points (Opara *et al.*, 2024; Morsidi, 2025).

2. Naïve Bayes

The Naive Bayesian Classifier is a popular and efficient classification technique frequently used in supervised learning. Naive Bayes models are often used in machine learning for their simplicity and efficiency (Islam, Bani and Ghosh, 2021; Guleria *et al.*, 2022; Çakir *et al.*, 2023). They fall under the category of classification methods that employ Bayes' Theorem. Bayes uses a unique data classification algorithm that makes the assumption that data attributes are independent of any particular class. This approach proves to be quite effective when dealing with datasets that have a lot of uncertainty and noise (Latt and Myint, 2015). Many tasks, such as text classification, spam filtering, sentiment analysis, and rating classification, gain significant advantages from the high-dimensional data produced through Bayesian classifications (Rácz, Bajusz and Héberger, 2021; Guleria *et al.*, 2022; Morsidi, 2025). Naive Bayes functions as a classifier, considering the features of the model as independent variables while also taking probabilities into account (Dehkordi, Manochehri and Aghazarian, 2023). In the illustration of Bayes' theorem, the characteristics of j classes is represented in an array of $y_i, j \in \{1, \dots, j\}$, while X is represented by $X = (X_1, X_2, \dots, X_m)$ comprising m

characteristics. The typical formulation assumption for expressing the Bayesian theorem is shown in Equation 2.

$$p(y_j|X) = \frac{p(X|y_j)p(y_j)}{p(X)} \quad (2)$$

Each characteristic has an independent impact on the predictions; a rare phenomenon found in data forecasting. Despite being widely proven to consist with lower probability of higher accuracies, recent studies have found evidence of Naïve Bayes' effectiveness while also suggesting difficulties with systematic improvement. By combining prior knowledge with training data, Bayesian learning algorithms can determine the posterior probability of a hypothesis (Saputra, Yudhana and Umar, 2022). The algorithm employs Bayes theorem throughout both training and prediction, rendering it a valuable resource for addressing classification problems and enhancing machine learning models (Çakir *et al.*, 2023). Bayesian learning algorithms leverage both prior knowledge and training data to calculate the posterior probability of a hypothesis.

This study referenced researcher's previous work (Morsidi, 2025), and attempt to configure the mechanism of both weighted polling KNN and Bayes classifier as a combined entity instead of working distinctively where the merging of these instance-based parametric learning is formulated in accordance with multiclass classification. A weighted voting ensemble approach has been based on the foundations of dry bean multiclass classifications, which have included several modifications to the original operational mechanisms for both KNN and Bayesian techniques. Prior to implementing the multiclass classification method, it is established that there is a necessity to consider factors such as optimal weights and computational efficiency within the research framework (Gültepe, 2022; Tavish Srivastava, 2024). Cross-validation, an evaluation metric in machine learning acting as balancer between training and testing data is deemed necessary in order to adjust the weightage polling from KNN and Bayes ensemble methods so that the dataset adheres to its distinct features (Pedroche *et al.*, 2020; Rosly *et al.*, 2024). This is stipulated to be able to lead to an increase in computational complexities when using instance-based learning for parametric and non-parametric techniques, as per the law of KNN and Naïve Bayes and observed from previous work (Morsidi, 2025). The following

lists down several variance in executing mechanism that is incorporated within the research scope.

- i. *Local Naive Bayes Training*: For a specific test case, the training set for the number of its k closest neighbours is determined using KNN, while the Naive Bayes classifier is also trained on this subset. The Naive Bayes model can be tailored to fit the local data distribution, resulting in improved performance from this approach. By taking advantage of both algorithms' benefits, the ensemble can improve classification accuracy, particularly in large datasets (Adi Pratama *et al.*, 2024).
- ii. *Weighted Voting Mechanism*: This combines the predictions of both KNN and the local Naive Bayes classifier, by which time each model's weight is assigned to its own prediction according to their confidence or past performance (Açikkar and Tokgöz, 2024; Morsidi, 2025); for example, if and especially in case Naive Bayes has historically performed better on some classes, it can be assumed that when predicting those classes this weight was greater than predicted because of global feature distributions, making the ensemble more resilient to changes in feature traits across classes.
- iii. *Probability-Based Integration*: By averaging the predicted probabilities from both classifiers, soft voting (minority class) replaces the hard and soft votes (majority class), which is used in weighted polling to capture the uncertainty in predictions (Rácz, Bajusz and Héberger, 2021; Bacevicius and Paulauskaite-Taraseviciene, 2023; Chabalala, Adam and Ali, 2023). This approach often yields better results when applied across multiple classes and permits an ensemble adjust to different class distributions and complexity, resulting in more balanced performance across all classes.

The KNN polling stage is used to determine the predominant class within the nearest neighbours for each data point needing classification. Distances to all points in the training set are calculated, and the nearest neighbours are determined, and the class label with the highest vote is

selected (Alsmadi and Almarashdeh, 2022; Tavish Srivastava, 2024). This repetitive pattern can consume a lot of computational power, especially when dealing with extensive datasets or data with many dimensions. Distance weighting and dimensionality reduction methods are used to enhance this process (Rosly *et al.*, 2024). The optimal k varies based on the data, with higher values decreasing noise but blurring class boundaries. Optimising hyper-parameters can help identify an optimal value for k . Weighted KNN and Weighted Naive Bayes are popular techniques in probability classification due to their improved predictive accuracy. Weighted KNN has more influence on the decision, leading to improved classification in complex datasets and lessening noise impact (Açikkar and Tokgöz, 2024; Morsidi, 2025). Naive Bayes with weighted models modifies probabilities to match the significance given to specific features, resulting in better performance in unbalanced and correlated data (Rácz, Bajusz and Héberger, 2021). The following table lists the implemented polling methods from the merging of both KNN and Bayes classifier.

Table 3 presents the essential weighted polling methods implemented in this study. The accuracy of the polling cycle can be greatly diminished by irrelevant or noisy attributes or inconsistent measurements of features. Naive Bayes employs probability computations instead of a core polling mechanism like KNN. Using Bayes' theoretical consideration, the underlying mechanism determines whether or not each test instance is associated with any given class and assumes that feature independence. This is the class denoted within the operational domain with the highest probability of best sample with greater accuracy.

Table 3. Weighted Polling Formulation for the Proposed Approach

Method	Formulation
Normal	$\sum_{n=0}^{samples} (Sum\ of\ total\ classification\ of\ nearest\ neighbour, n) + 1$
Weighted	$\sum_{n=0}^{samples} (Sum\ of\ total\ classification\ of\ nearest\ neighbour, n) + (samples - n)$
Weight-Squared	$\sum_{n=0}^{samples} (Sum\ of\ total\ classification\ of\ nearest\ neighbour, n) + (samples - n)^2$

D. Detailed Weighted Polling Mechanism

The proposed weighted polling ensemble integrates KNN and Naïve Bayes through a structured multi-stage process:

1. Weight Determination Strategy

The weights attached to each classifier are established via k-fold cross-validation ($k=5$) on the training data in this case. KNN and Naïve Bayes are trained for each fold separately, and their respective accuracies (Acc_{KNN} and Acc_{NB}) are reported. The last weights are calculated as:

$$w_{KNN} = Acc_{KNN} / (Acc_{KNN} + Acc_{NB})$$

$$w_{NB} = Acc_{NB} / (Acc_{KNN} + Acc_{NB})$$

Thus, it is guaranteed that decision makers with higher validation scores will impact the end prediction more. Weights are determined through cross-validation, and then they are used the same way during the testing phase.

2. Local Naïve Bayes Training Process

Each test instance x_{test} is treated as follows:

- KNN finds the K-Nearest Neighbors from the training set with the help of the Euclidean distance
- A local set S_{local} is formed that includes only those k neighbors along with their labels
- A Naïve Bayes model NB_{local} is developed based solely on S_{local}
- This model that has been trained locally predicts the class probability $P_{NB_{local}}(c|x_{test})$ for every class c

This technique of local training makes the Naïve Bayes component flexible enough to fit the local data distribution surrounding each test point, rather than being dependent on global class statistics alone.

3. Prediction Integration

Both classifiers produce probability distributions of classes for every single test instance:

- KNN presents $P_{KNN}(C|x_{test})$ which is grounded on the class occurrence of nearest k neighbours
- Local NB rolls out $P_{NB_{local}}(c|x_{test})$ from the locally trained model

The weighted ensemble prediction is computed as:

$$P_{ensemble}(C|x_{test}) = w_{KNN} \times P_{KNN}(C|x_{test}) + w_{NB} \times P_{NB_{local}}(C|x_{test})$$

The final predicted class is:

$$\hat{y} = \operatorname{argmax}_c P_{ensemble}(C|x_{test})$$

4. Distinction from Standard Voting Ensembles

There are three main differences between our method and the classic voting ensembles:

- **Adaptive Local Modeling:** The Naïve Bayes part is retrained for every test instance according to its K-Nearest Neighbors rather than applying one global model
- **Performance-Based Weighting:** The weights come from the empirical validation performance rather than being fixed equally or arbitrarily assigned
- **Probability-Level Fusion:** The experiment merge probability distributions instead of hard class labels, thus maintaining prediction uncertainty and allowing more sophisticated decision-making.

IV. RESULTS & DISCUSSION

Class distribution plot for the referenced dataset demonstrated a significant disparity within the dataset domain, where DERMASON class has 4500 samples, and the BOMBAY class consisted of only 500 sample. Violin plots are also applied for visualising multiclass classification tasks in the Dry Bean dataset show the complete distribution of numerical features across various categories in order to effectively representing multimodal distributions with multiple peaks consisting of feature distributions across distinct bean types.

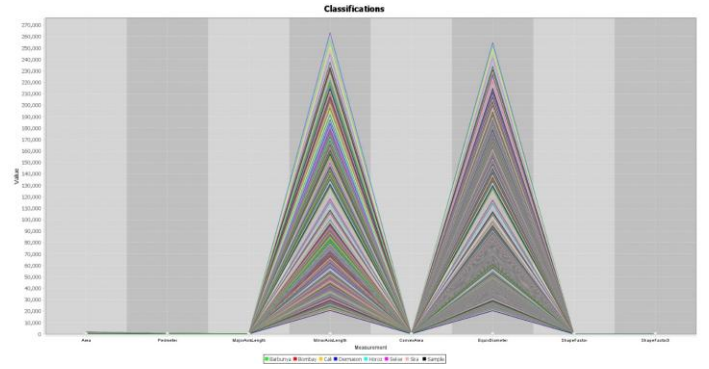


Figure 2. Results of Multiclass Classification based on 7 Distinct Dry Bean Types from Combination of KNN and Naïve Bayes Weighted Polling Measure

Figure 2 displays a parallel coordinates visualisation emphasising feature value ranges and inter-class variability across all eight morphological dimensions such as BARBUNYA, BOMBAY, CALI, DERMASON, HOROZ, SEKER, and SIRA, across several morphological features, including *Area*, *Perimeter*, *MajorAxisLength*, *MinorAxisLength*, *ConvexArea*, *EquivDiameter*, *ShapeFactor*, and *ShapeFactor3*. Each vertical axis denotes a unique attribute, and every colored line symbolises a sample from a particular class. The graph indicates that features like *MajorAxisLength*, *ConvexArea*, and *EquivDiameter* exhibit a broader range of values, implying greater variability among classes, which may enhance their effectiveness in classification tasks. On the other hand, features such as *ShapeFactor* and *ShapeFactor3* show minor differences and considerable overlap among classes, suggesting a lower capacity to distinguish between them. By utilising the weighted polling measure, which involves KNN instance learning and Nave Bayes probability classifier, it is possible to determine the best specimen representing each cluster creation, with the selected sample representing the overall prime quality dry bean for entire domain.

To verify patterns, anomalies, and relationships within the data cluster, the specimen domain is tested using several exploratory data analysis (EDA) before the proposed probability classifier approach is imposed on the referenced dataset. This process is essential for developing a general hypothesis about the structural decomposition of the entire multiclass dataset, including identification of class imbalances, insights into the information contained in the database, as well as comparisons between different groups

under discrete circumstances that affect data variability and influence the percentage of accuracy rate.

1. Dry Bean Representation in Scatter Plot with Basic KNN Visualisation

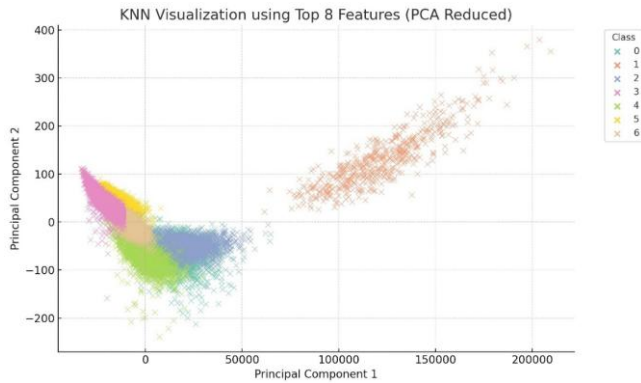


Figure 3. Scatter Plot Visualization using basic KNN clustering on the referenced dry bean dataset

Figure 3 demonstrates the PCA-based scatter plot highlighting global class separability in reduced 2D space, demonstrating clear separation of BOMBAY class. For analysing intrinsic multiclass clusters, PCA is utilised in this scatter plot to extract a KNN visualisation of seven distinctive bean classes (labelled 0 to 6) using the top eight features. Points are used to represent a sample, and their location is determined by the first two principal components, which capture most of the dataset's significant variation. The orange-coloured Class 1 is distinctly different from all other classes along Principal Component 1, indicating strong discriminative traits. Conversely, Classes 0, 2, 3, 4, 5, and 6 are found in clusters closer to their origins, implying some shared characteristics among them that could lead to potential classification difficulties.

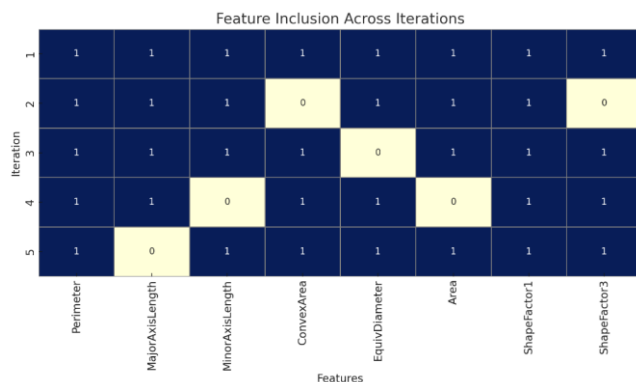


Figure 4. Correlation Heatmap Visualizing Inclusion of Certain Feature Traits in Determining the Predictive Accuracy of Dry Bean Class

The heatmaps visually exhibited the effects of feature selection on predicting accuracy for all five iterations. The feature inclusion heatmap indicates which features were included (✓) and excluded (X) in each iteration. The features *Perimeter* and *ShapeFactor1* were included in each of the five iterations but are naturally model features. Features that had more variation across each iteration included *ConvexArea*, *ShapeFactor3*, and *EquivDiameter*. Each of these features demonstrated an ability to drive increased accuracy on their own. The accuracy heatmap illustrated the impact the features in each iteration had on the performance of the model. Iteration 3, at 88.58%, represents the combination of features the model was able to most optimally use there were iterations two and five equating to the lowest accuracy (85.67%), in which iterations two and five both excluded any features included and consistently included from either iteration 1, 4, or 3 and accuracy was maximised. Collectively, these heatmaps illustrated and displayed how the addition of features directly affects the classification accuracy of the model and the experimental goal behind feature selection is twofold, identifying the efficient use of the proper features with predictive power but also to add a more intentional strategic layer to selective feature inclusion; all while maximising accuracy in predictive models.

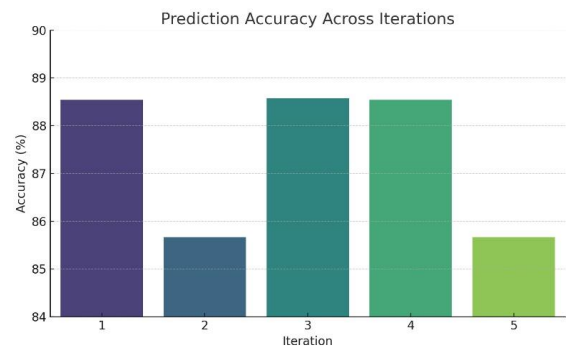


Figure 5. Results of Prediction Accuracy for Entire Cluster of Dry Beans Across 5 Iterations

The predictive classification accuracy across iterations ranged from 85.67%-88.58%. All iterations had moderate accuracy variation based on features selected for inclusion. The iteration 3 produced the highest accuracy (88.58%) indicating it included the correct sets of features. Iteration 1 and 4 were very close to each other (88.54%) which felt consistent in performance. Contrasting this, Iteration 2, and Iteration 5 produced the least accurate results (85.67%) suggesting that removing features in these iterations limited

the model's predictive capability. Overall, the results of the iterations show the model's sensitivity to feature selection, which shows the potential improvement of performance through further refinement of input features.

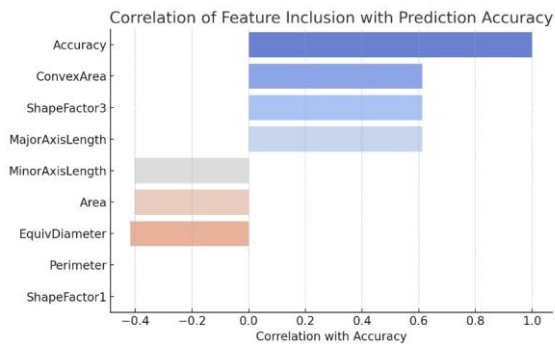


Figure 6. Correlation Analysis of Distinctive Dry Bean Category according to Major Feature Traits

The results from the feature ablation table seem to indicate that some features considerably influence the accuracy of the dry bean classification model. The features *ConvexArea*, *ShapeFactor3*, and *MajorAxisLength* all show some moderate positive correlation with prediction accuracy, supporting the model performance improvement with the use of those features. The features *MinorAxisLength*, *Area* and *EquivDiameter* show a moderate negative correlation with accuracy, indicating that these three features slightly decreased predictive performance on average when included. It should be noted that *Perimeter* and *ShapeFactor1* were included in all iterations, thus no correlation analysis could be performed. However, having those two features in every iteration also supports the assumption that they were baseline features. In all, the iteration which provided the highest accuracy (88.58%) would potentially give the best features to use for an optimized classification effort and demonstrated the value of feature selection in predicting dry bean classification outcomes.

Table 4 in the Appendix presents the list of feature traits and number of iterations used to identify a prime candidate from each phase, along with an emphasis on analysis for surface area (*ConvexArea*) and overall diameter (*Area*). Among these two features, which are known to differ greatly across classes, they were particularly influential in explaining the class segregation. The main components are believed to have been influenced by *ConvexArea* and *Area*, as evidenced by the distinct separation of Class 1 in the PCA-based KNN

visualisation. Unique clustering supports this, where the feature can be imposed to determine whether certain parameters influence classification results, either by adding them or removing them through yes/no combinations. By using both *ConvexArea* and *Area*, iterations were able to produce more distinct cluster distinctions, which is why they are frequently used in choosing better samples.

2. Bar Chart of the Dry Bean Dataset by Class Hierarchies

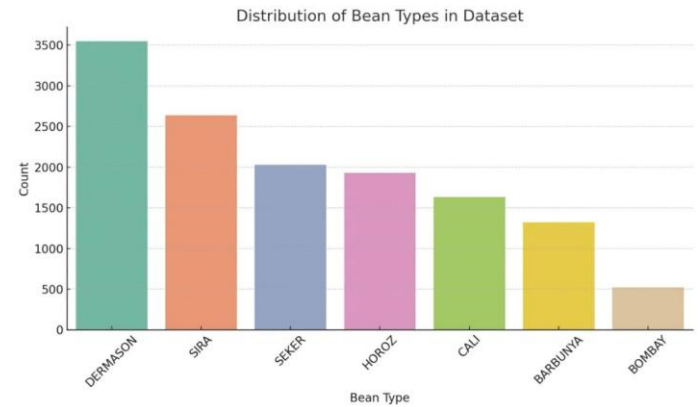


Figure 7. The class hierarchies of bean type distributions within the entire dataset

The bar chart displays the amount of samples constitution representative for each bean type in the dataset. DERMASON has the highest number of samples among the seven classes, with over 3,500 entries, followed by SIRA and SEKER, both with more than 2,000 entries. Although HOROZ, CALI and BARBUNYA have moderate representation, BOMBAY has fewer than 1,000 samples and is of very low proportions. This class imbalance has important consequences for model training and assessment: unless appropriate methods such as class weighting or resampling techniques are used, classifiers may bias towards the majority classes like DERMASON. This proposed experiment is compatible with being tested and compared with the unaltered version of KNN instance learning approach. Also, underrepresented groups such as BOMBAY could encounter reduced training examples and consequently lower accuracy or higher misclassification rates. DERMASON is the most frequently occurring bean type in the dataset, with SIRA and SEKER being the next popular choice. This point exhibited the consistent trend for the multiclass dataset based on significant difference between sample quantities, however BARBUNYA and BOMBAY are distinct exceptions from this

trend in the sense of lower quantities. This may influence the decrease in performance for training models due to an uneven distribution, leading to fewer accurate predictions for minority classes and therefore biased towards majority classes. BOMBAY and other multiclass individuals may exhibit high overall accuracy but poor performance on classes with fewer samples. This further testifies the need for using balancing techniques such as data augmentation, stratified sampling, or class-weighted models during training. Moreover, the perceived inequality may affect the applicability of performance metrics like accuracy rate.

3. Violin Plot of the Dry Bean Dataset by Class Hierarchy

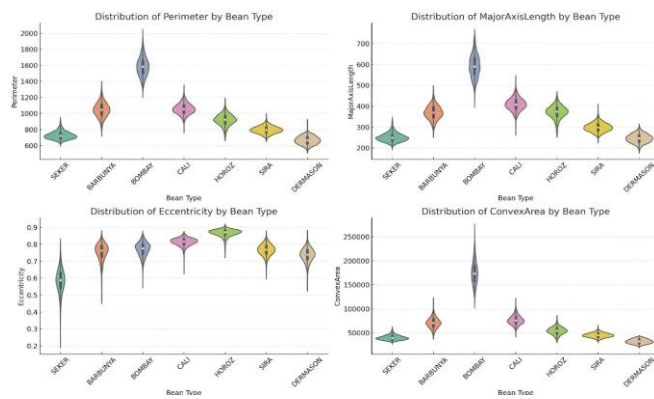


Figure 8. Violin Plot of the Dry Bean Cluster analysed on *Perimeter*, *MajorAxisLength*, *Eccentricity*, and *ConvexArea*

In the violin plot for Figure 8, the top three types of beans (BOMBAY, BARBUNYA and CALI) and the bottom three (SEKER, SIRA and DERMASON) are shown to have different distributions of key morphological traits: *Perimeter*, *MajorAxisLength*, *Eccentricity* and *ConvexArea*, based on size-related features. BOMBAY beans have wide distributions and the highest values for perimeter, major axis length, and convex area consistent with earlier observations of within-type morphological variability. The metrics for these features are elevated, but the spreads are relatively narrower in BARBUNYA and CALI, indicating more uniform morphology. On the other hand, the beans produced by SEKER, SIRA, and DERMASON are smaller and more compact in all three size ranges, suggesting that they have limited variability due to their relatively homogeneous morphology. CALI, BARBUNYA, BOMBAY, and SIRA all exhibit eccentricity values that suggest a longer bean shape, while SEKER shows the greatest variance in eccentricities,

indicating shape heterogeneity. DERMASON exhibits a tighter distribution with lower eccentricity values, which is consistent with rounded shapes.

The violin plot depict the distributions of key morphological traits—*Perimeter*, *MajorAxisLength*, *Eccentricity*, and *ConvexArea* among the top three (BOMBAY, BARBUNYA, CALI) and bottom three (SEKER, SIRA, DERMASON) bean types based on size-related characteristics. Consistent with prior observations, BOMBAY beans exhibit the highest values for perimeter, major axis length, and convex area, accompanied by wide distributions, indicating considerable within-type morphological variability. BARBUNYA and CALI also demonstrate elevated metrics for these features but with relatively narrower spreads, reflecting more uniform morphology. In contrast, SEKER, SIRA, and DERMASON beans are characterised by smaller and more compact distributions across all three size parameters, highlighting their morphological consistency and limited variability. Regarding shape, eccentricity values are highest for CALI, followed closely by BARBUNYA, BOMBAY, and SIRA, suggesting a more elongated bean shape, while SEKER again displays the greatest variance in eccentricity, indicating shape heterogeneity. DERMASON shows a tighter distribution with lower eccentricity values, consistent with a more rounded form. Overall, the plots underscore clear morphological stratification, with BOMBAY, BARBUNYA, and CALI occupying the upper bounds of size and elongation, whereas SEKER, SIRA, and DERMASON exhibit more compact and uniform morphological profiles.

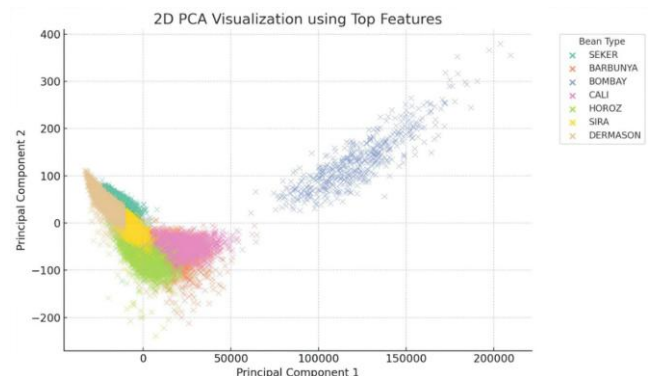


Figure 9. A 2D PCA Analysis on Dry Bean Dataset Composition based on Density & Scattered Populative Disposition of Individual Samples

The 2D PCA plot for this dry bean dataset shown in Figure 9 illustrates density-enhanced PCA visualisation revealing within-class compactness and overlap patterns among the seven bean varieties. This clearly shows that these BOMBAY beans (light blue) form a unique and tightly grouped cluster on the far right, indicated and distinctly separated along the first principal component. This could be attributed to prominent physical characteristics like size or form. Conversely, SEKER, BARBUNYA, CALI, HOROZ, SIRA, and DERMASON are tightly grouped in their central areas, indicating more nuanced differences and shared characteristics. In this overlapping cluster, the internal arrangements of SIRA and its closely related domain group HOROZ are more compact, indicating more consistent internal traits despite their visual integration with other types. CALI beans are positioned slightly to the right, suggesting some degree of preference. The fan-shaped configuration of points denotes correlated traits, with Principal Component 1 accounting for much of the variation and serving as the primary factor for separation.

4. Performance Interpretation Considering Dataset Characteristics

The accuracy of 88.58% should be seen in the light of the class distribution as presented in the dataset (Figure 7). The distribution shows that DERMASON represented 26% and BOMBAY only 4% of the total samples; thus, it is difficult to gain full insight into per-class performance from the overall accuracy alone. The accuracy obtained, being relatively high, can be taken as a sign of successful classification beyond the simple major-class prediction; yet, without per-class metrics like precision and recall for each bean variety, it is impossible to make a definitive conclusion whether the minority classes (especially BOMBAY with only 522 samples) have comparable performance as the majority classes do. Morphological visualisations (Figures 8-10) display clear feature-space separation for some classes (especially BOMBAY), which might be the reason for the correct classification of such classes even when limited training data is available. On the other hand, the considerable overlap seen among SEKER, SIRA, and DERMASON implies that these classes might be subject to higher misclassification rates, which in turn are concealed by the overall accuracy metric.

V. CONCLUSION AND RECOMMENDATION

This investigation highlights the effectiveness of a combo probability-based classification approach in the field of agriculture, demonstrating the use of a weighted voting ensemble of K-Nearest Neighbors and Naïve Bayes classifiers for dry bean variety identification over seven classes. The method proposed resulted in classification accuracy of 85.67% to 88.58% depending on the dataset, the KNN's instance-based learning and neighbourhood-specific distributions adapted locally by Naïve Bayes models complementing each other in strengths. The application of dimensionality reduction and visualisation techniques (PCA, violin plots, parallel coordinates) brought to light important factors regarding the separation of classes, including the particularly clear morphological separation of BOMBAY beans. The study also points out some limitations in methods that can be considered as great opportunities for further studies. First, overall accuracy was the only metric used throughout the study; considering the large difference between classes (BOMBAY: 522 samples while DERMASON: 3,546 samples), the performance of the model on minority classes should be evaluated through a thorough assessment including macro-averaged F1-score, balanced accuracy, per-class precision/recall metrics, and confusion matrix analysis. Second, the experimental procedure would lose nothing and gain in rigor with more extensive documentation such as the different datasets used, partitioning techniques employed, cross-validation methods adopted, feature preprocessing strategies, hyperparameter tuning done, and random seed control for ensuring reproducibility. The value of weighted polling would be more convincing if it is directly compared to baseline methods (standard KNN, standard Naïve Bayes, unweighted voting, Random Forest, Gradient Boosting). Quantitative analysis of computational efficiency (training time, prediction latency, memory consumption) should provide empirical evidence to support the claim of advantages over complex ensemble methods. In this work's contextual output, future research directions include dynamic weighting schemes for real-time classification, transfer learning across agricultural crops, integration with computer vision pipelines, and the development of mechanisms to handle changing class distributions. The limitations addressed by hybrid ensemble approaches will

strengthen both the scientific completeness and practical applicability of such approaches for agricultural classification tasks. Moreover, this study provides invaluable insights into probability-weighted classification for multiclass problems that are not only applicable in agriculture but also in related domains requiring interpretable and efficient automated decision systems.

VII. REFERENCES

- Açıkkar, M. and Tokgöz, S. (2024) 'An improved KNN classifier based on a novel weighted voting function and adaptive k-value selection', *Neural Computing and Applications*, 36(8), pp. 4027–4045. Available at: <https://doi.org/10.1007/s00521-023-09272-8>.
- Adi Pratama, I.P. *et al.* (2024) 'Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties', *Indonesian Journal of Data and Science*, 5(1), pp. 23–29. Available at: <https://doi.org/10.56705/ijodas.v5i1.124>.
- Alsmadi, M.K. and Almarashdeh, I. (2022) 'A survey on fish classification techniques', *Journal of King Saud University - Computer and Information Sciences*, 34(5), pp. 1625–1638. Available at: <https://doi.org/10.1016/j.jksuci.2020.07.005>.
- Bacevicius, M. and Paulauskaite-Taraseviciene, A. (2023) 'Machine Learning Algorithms for Raw and Unbalanced Intrusion Detection Data in a Multi-Class Classification Problem', *Applied Sciences (Switzerland)*, 13(12). Available at: <https://doi.org/10.3390/app13127328>.
- Çakir, M. *et al.* (2023) 'Accuracy assessment of RFerns, NB, SVM, and kNN machine learning classifiers in aquaculture', *Journal of King Saud University - Science*, 35(6). Available at: <https://doi.org/10.1016/j.jksus.2023.102754>.
- Chabalala, Y., Adam, E. and Ali, K.A. (2023) 'Exploring the Effect of Balanced and Imbalanced Multi-Class Distribution Data and Sampling Techniques on Fruit-Tree Crop Classification Using Different Machine Learning Classifiers', *Geomatics*, 3(1), pp. 70–92. Available at: <https://doi.org/10.3390/geomatics3010004>.
- Dehkordi, I.F., Manochehri, K. and Aghazarian, V. (2023) 'Internet of Things (IoT) Intrusion Detection by Machine Learning (ML): A Review', *Asia-Pacific Journal of Information Technology and Multimedia*, 12(1), pp. 13–38. Available at: <https://doi.org/dx.doi.org/10.17576/apjitm-2023-1201-02>.
- Demir, K. and Yaman, O. (2024) 'A HOG Feature Extractor

VI. ACKNOWLEDGEMENT

The author would like to acknowledge Universiti Pendidikan Sultan Idris, which supported this research and subsequent publications. Further appreciation is extended to the peers who reviewed and provided positive feedback regarding the article. There is no financial support from any funding agency in either the public, private or not-for-profit sectors for this research.

- and KNN-Based Method for Underwater Image Classification', *Firat University Journal of Experimental and Computational Engineering*, 3(1), pp. 1–10. Available at: <https://doi.org/10.62520/fujece.1443818>.
- Fondón, I. *et al.* (2018) 'Automatic classification of tissue malignancy for breast carcinoma diagnosis', *Computers in Biology and Medicine*, 96(December 2017), pp. 41–51. Available at: <https://doi.org/10.1016/j.compbiomed.2018.03.003>.
- Guleria, K. *et al.* (2022) 'Early prediction of hypothyroidism and multiclass classification using predictive machine learning and deep learning', *Measurement: Sensors*, 24(August), p. 100482. Available at: <https://doi.org/10.1016/j.measen.2022.100482>.
- Gültepe, Y. (2022) 'Analysis of Alburnus tarichi population by machine learning classification methods for sustainable fisheries', *SLAS Technology*, 27(4), pp. 261–266. Available at: <https://doi.org/10.1016/j.slant.2022.03.005>.
- Islam, S.M.M., Bani, S.I. and Ghosh, R. (2021) 'Content-based Fish Classification Using Combination of Machine Learning Methods', *International Journal of Information Technology and Computer Science*, 13(1), pp. 62–68. Available at: <https://doi.org/10.5815/ijitcs.2021.01.05>.
- Javed Mehedi Shamrat, F.M. *et al.* (2022) 'LungNet22: A Fine-Tuned Model for Multiclass Classification and Prediction of Lung Disease Using X-ray Images', *Journal of Personalized Medicine*, 12(5). Available at: <https://doi.org/10.3390/jpm12050680>.
- Koklu, M. and Ozkan, I.A. (2020) 'Multiclass classification of dry beans using computer vision and machine learning techniques', *Computers and Electronics in Agriculture*, 174, p. 105507. Available at: <https://doi.org/https://doi.org/10.1016/j.compag.2020.105507>.
- Kuru, K. and Khan, W. (2018) 'Novel hybrid object-based non-parametric clustering approach for grouping similar objects in specific visual domains', *Applied Soft Computing*

- Journal*, 62, pp. 667–701. Available at: <https://doi.org/10.1016/j.asoc.2017.11.007>.
- Latt, S.S. and Myint, M. (2015) ‘Classification of Fish based on Naive Bayesian Classifier’, in *LISS 2013*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 585–589. Available at: https://doi.org/10.1007/978-3-642-40660-7_86.
- Moral, P. Del, Nowaczyk, S. and Pashami, S. (2022) ‘Why Is Multiclass Classification Hard?’, *IEEE Access*, 10(June), pp. 80448–80462. Available at: <https://doi.org/10.1109/ACCESS.2022.3192514>.
- Morsidi, F. (2025) ‘Hybrid Learning for Bottlenose Dolphin Motion Approximation Using Weighted Polling KNN and Bayes Mechanism’, *International Journal of Software Engineering and Computer Systems*, 10(2), pp. 173–186. Available at: <https://doi.org/https://doi.org/10.15282/ijsecs.10.2.2024.14.0132>.
- Opara, I.K. *et al.* (2024) ‘Machine Learning Application in Horticulture and Prospects for Predicting Fresh Produce Losses and Waste: A Review’, *Plants*, 13(9), pp. 1–21. Available at: <https://doi.org/10.3390/plants13091200>.
- Pedroche, D.S. *et al.* (2020) ‘Architecture for trajectory-based fishing ship classification with AIS data’, *Sensors (Switzerland)*, 20(13), pp. 1–21. Available at: <https://doi.org/10.3390/s20133782>.
- RÁCZ, A., Bajusz, D. and Héberger, K. (2021) ‘Effect of dataset size and train/test split ratios in qsar/qspr multiclass classification’, *Molecules*, 26(4), pp. 1–16. Available at: <https://doi.org/10.3390/molecules26041111>.
- Rosly, R. *et al.* (2024) ‘Multi-classifier models to improve the accuracy of fish landing application’, *International Journal of Advanced Technology and Engineering Exploration*, 11(111), pp. 145–159. Available at: <https://doi.org/10.19101/IJATEE.2023.10102060>.
- Salaudin Khan, M. *et al.* (2023) ‘Comparison of multiclass classification techniques using dry bean dataset’, *International Journal of Cognitive Computing in Engineering*, 4(January), pp. 6–20. Available at: <https://doi.org/10.1016/j.ijcce.2023.01.002>.
- Saputra, S., Yudhana, A. and Umar, R. (2022) ‘Implementation of Naïve Bayes for Fish Freshness Identification Based on Image Processing’, *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, 6(3), pp. 412–420. Available at: <https://doi.org/10.29207/resti.v6i3.4062>.
- Sarki, R. *et al.* (2022) ‘Convolutional Neural Network for Multi-class Classification of Diabetic Eye Disease’, *EAI Endorsed Transactions on Scalable Information Systems*, 9(4). Available at: <https://doi.org/10.4108/eai.16-12-2021.172436>.
- Sebak, M. and Kacsuk, Z. (2021) ‘The Multiclass Classification of Newspaper Articles with Machine Learning: The Hybrid Binary Snowball Approach’, *Political Analysis*, 29(2), pp. 236–249. Available at: <https://doi.org/10.1017/pan.2020.27>.
- Shamrat, F.J.M. *et al.* (2023) ‘High-precision multiclass classification of lung disease through customized MobileNetV2 from chest X-ray images’, *Computers in Biology and Medicine*, 155(June 2022), p. 106646. Available at: <https://doi.org/10.1016/j.compbimed.2023.106646>.
- Suryawanshi, P. and Dutta, P. (2022) ‘Optimization models for supply chains under risk, uncertainty, and resilience: A state-of-the-art review and future research directions’, *Transportation Research Part E: Logistics and Transportation Review*, 157(December 2021), p. 102553. Available at: <https://doi.org/10.1016/j.tre.2021.102553>.
- Tavish Srivastava (2024) ‘Guide to K-Nearest Neighbors Algorithm in Machine Learning’, <https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/>, pp. 1–20.
- Taylor, R. (2023) ‘Machine Learning Techniques for Fish Breeding Decision Making’, *2023: Wellington Faculty of Engineering Symposium*, pp. 1–12. Available at: <https://ojs.victoria.ac.nz/wfes/article/view/8422>.
- Vidarthi, A. *et al.* (2022) ‘Machine Learning Assisted Methodology for Multiclass Classification of Malignant Brain Tumors’, *IEEE Access*, 10, pp. 50624–50640. Available at: <https://doi.org/10.1109/ACCESS.2022.3172303>.
- Wang, K., Muthukumar, V. and Thrampoulidis, C. (2023) ‘Benign Overfitting in Multiclass Classification: All Roads Lead to Interpolation’, *IEEE Transactions on Information Theory*, 69(12), pp. 7909–7952. Available at: <https://doi.org/10.1109/TIT.2023.3320098>.
- Yilmaz, A.E. and Demirhan, H. (2023) ‘Weighted kappa measures for ordinal multi-class classification performance’, *Applied Soft Computing*, 134, p. 110020. Available at: <https://doi.org/10.1016/j.asoc.2023.110020>.

Appendix

Table 4. Several Iterations of Prediction on the Best Dry Bean Class using the Proposed Method

Iteration	Perimeter	MajorAxisLength	MinorAxisLength	ConvexArea	EquivDiameter	Area	ShapeFactor1	ShapeFactor3	Predicted Class	%
1	✓	✓	✓	✓	✓	✓	✓	✓	HOROZ	88.54
	✓	✓	✓	✓	✓	✗	✓	✓	HOROZ	85.67
	✓	✓	✓	✗	✓	✓	✓	✓	HOROZ	88.58
2	✓	✓	✓	✓	✓	✓	✓	✓	HOROZ	88.54
	✓	✓	✓	✓	✓	✗	✓	✓	HOROZ	85.67
	✓	✓	✓	✗	✓	✓	✓	✓	HOROZ	88.58
3	✓	✓	✓	✓	✓	✓	✓	✓	HOROZ	88.54
	✓	✓	✓	✓	✓	✗	✓	✓	HOROZ	85.67
	✓	✓	✓	✗	✓	✓	✓	✓	HOROZ	88.58
4	✓	✓	✓	✓	✓	✓	✓	✓	HOROZ	88.54
	✓	✓	✓	✓	✓	✗	✓	✓	HOROZ	85.67
	✓	✓	✓	✗	✓	✓	✓	✓	HOROZ	88.58
5	✓	✓	✓	✓	✓	✓	✓	✓	HOROZ	88.54
	✓	✓	✓	✓	✓	✗	✓	✓	HOROZ	85.67
	✓	✓	✓	✗	✓	✓	✓	✓	HOROZ	88.58