Technium Vol. 22, pp.14-32 (2024)
ISSN: 2668-778X
www.techniumscience.com

Naive Bayes Classifier in Grading Carabao Mangoes

Gary P. Guillergan [1*], Reymund L. Sabay [2], Joel M. Bual [3]
[1] *Northern Iloilo State University, Batad, Iloilo, Philippines*
[2,3] *University of Negros Occidental-Recoletos, Bacolod City, Philippines*
[*] *Corresponding author: matrix5016@yahoo.com*

Abstract

This study explores machine learning's potential to classify carabao mangoes, a key Philippine export, into four grades based on size: A (large), B (medium), C (small), and R (reject). It introduces a Naïve Bayes classification model that uses image processing to extract features for grading. The goal is to create a consistent grading system to enhance export efficiency and benefit local farmers. The research aims to validate the Naïve Bayes model's accuracy using size, weight, area, and spot ratio. It employs a quantitative, experimental design, manipulating image processing techniques to gauge their impact on classification accuracy. The results show the Naïve Bayes model achieved 95% accuracy, effectively distinguishing large and reject mangoes. It performed well for medium and small mangoes, with a 7% error rate between these classes. This indicates the model's potential for quality control and sorting, though further refinement is needed to better differentiate between medium and small sizes. In conclusion, the study presents an image processing and Naïve Bayes-based method to classify carabao mangoes by size. The model's high accuracy suggests its effectiveness and potential for automating mango classification, which could significantly aid the Philippine mango industry. Further performance assessment was conducted using a confusion matrix. The research highlights the promise of this approach for efficient mango grading.

Keywords: Naïve Bayes classifier, carabao mango, quantitative-experimental, Philippines

1.0 INTRODUCTION

Due to its distinctive flavor and sweetness, the carabao mango stands as the predominant and exported mango type in Philippines, elevating the country's recognition worldwide [1]. Mangoes serve as the main revenue generator for the agricultural sector in the country and represent one of its crucial export commodities [2]. Renowned as the world's sweetest mango, the carabao variety enjoys widespread popularity [3]. Meanwhile, the farmers must classify these mangoes as quality to meet the export standards [4]. Turning to innovation in Philippine agriculture, artificial intelligence (AI) is playing a growing role. While carabao mangoes have long been a cornerstone of the Philippine economy, AI offers new ways to optimize their cultivation and ensure export quality [5].

Furthermore, AI trains machines to achieve tasks that typically require human intelligence. Its objective is to develop computer software and hardware systems capable of emulating human processes that demonstrate traits linked with their intelligence [6]. Interestingly, machine learning refers to computers as opposed to humans. Here, the concept of learning is the same whether the learner is human or machine. While machine learning holds promise for broader applications, the current forefront of the field often finds itself confined to addressing individual problem scenarios [7]. Meanwhile, in supervised learning, a machine learning algorithm is furnished with a significant dataset comprising sample inputs for the intended output or classification, often curated with input from a domain expert. The algorithm's aim is to identify patterns within the data and establish comprehensive rules for linking inputs to their respective classes or events [8].

Meanwhile, a Naive Bayesian model is a probabilistic classifier that uses the Bayes hypothesis and naive presumptions about characteristic freedom [9]. It assumes that a particular

characteristic's value is free from other elements' qualities, also known as conditional independence [10]. Despite its naive assumption, Bayesian classifiers have proven helpful in complex real-world conditions. They are highly extensible and require specific parameters in the number of variables, making learning difficult. Naive Bayes classifiers are a standard practice method but a family of calculations [11].

Additionally, Oliver [12] believed that mango is the third largest fruit crop in Philippines, close to banana and pineapple. Its importance is derived from more than just the export side. There are three well-known varieties of Philippine mangoes: carabao, pico, and katchamita [13]. However, carabao mango is the dominant variety widely grown nationwide and exported. Pauly and Sankar [14] perceived that one factor influencing the mangoes export is the categorization and evaluation procedures after the harvest. Categorization is classifying mangoes into distinctive groups based on their diversity while evaluation is the quality-based classification process.

Based on the work review, the researchers use different procedures for fruit grading and sorting. One approach that works well for a fruit or vegetable may not work well for another [15]. Techniques to classify relied on predetermined models, with some stage-by-stage fruit classification completed. Fruit classification remains challenging due to the significant number of characteristics such as form, size, color, texture, and intensity [16]. Interestingly, there is a literature gap on the need for consistent and globally approved mango categorization methods. Studies on these areas can lead to developing new and more efficient ways of grading or sorting mangoes and improving its accuracy. New tools and systems for grading and sorting may have social and economic implications for various stakeholders.

Thus, this study developed an accurate Naïve Bayes classification model to classify the grading of carabao mango fruits based on size, weight, area, and spot ratio. The Naïve Bayes Classifier was used to create a classification method capable of classifying mangoes into categories of grade A (large), grade B (medium), grade C (small), and grade R (reject). The research concentrated on the carabao mango variety. Likewise, it investigated the difference in the accuracy of the Naïve Bayes classification model relative to the demographics.

2.0 FRAMEWORK OF THE STUDY

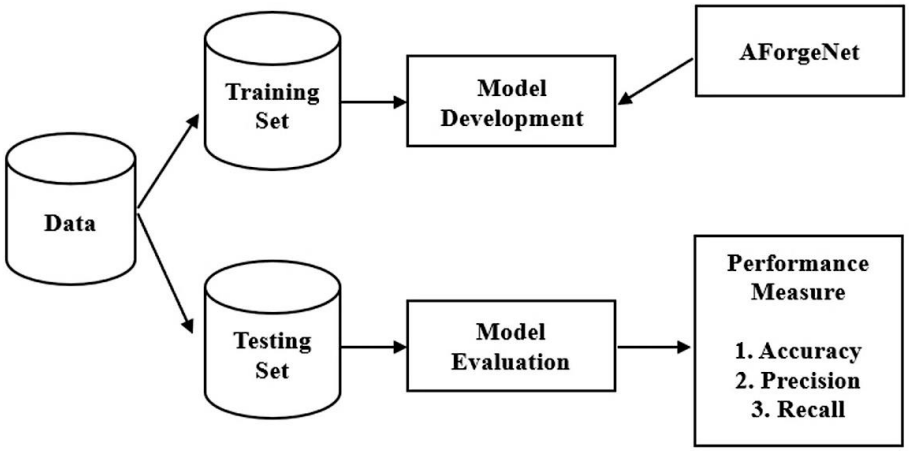
Physical appearance is an essential carabao mango characteristic and its quality. The size, weight, area, and spot ratio are extracted to classify this variety. Meanwhile, machine learning includes various techniques, including supervised learning and tasks like classification. One specific classification method is Naïve Bayes, which leverages probability concepts to make predictions. Naïve Bayes classifier is based on Bayesian theorem with the naïve assumption of independence between each feature pair. Its key advantage is the conditional independence assumption, which helps obtain rapid classification and probabilistic hypotheses. It serves as a foundation for numerous machine learning and data mining approaches in creating predictive models. This technique is beneficial in scenarios where input dimensionality is high [17]. Also, it’s a robust machine learning methodology that delivers quick and accurate classifications despite its simplistic independence.

The Naïve Bayes classifier relies on a naïve dependent feature model, assuming that the occurrence or absence of 1 attribute is independent of the occurrence or absence of another attribute. One advantage is its minimal training data requirement for classification. It also assumes attribute independence, allowing for efficient estimation of document class membership by estimating probabilities based on the unrelatedness of 1 attribute to another

feature. Naïve Bayes is mainly used when the naïve puts are high. The predictable state shows the probability of each input attribute. It uses training data to estimate parameters for classification, determining the probabilities of attributes in different classes. The Naïve Bayes classifiers, akin to linear models, form a family of classifiers known for their expediency in training. These models excel in efficiency because they ascertain parameters by scrutinizing each feature independently and gathering straightforward statistics per class. They exhibit swift training and prediction capabilities, with a straightforward training process. Naïve Bayes models serve as solid foundational models, commonly applied to large datasets [18].

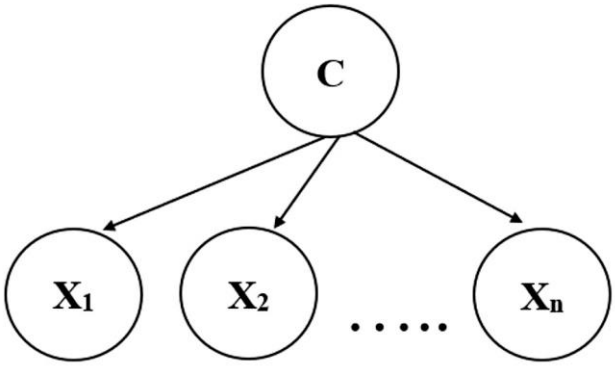
The classification using the Naïve Bayes model involves 2 distinct phases: a learning phase and an evaluation phase. During the learning phase, the classifier trains the model using a provided dataset, while in the evaluation, it assesses the classifier's performance. Specifically, evaluation is based on diverse parameters, including accuracy, error rate, and recall [19].

Figure 1. *Naïve Bayes Model*



Naïve Bayes represents the simplest form of a Bayesian network, assuming that all attributes are independent given the class variable's value, known as conditional independence. However, this conditional independence assumption is seldom accurate in many scenarios. A straightforward approach to controlling Naïve Bayes' limitation is increasing its structure to explicitly represent the dependencies among attributes [20]. Naïve Bayes assumes conditional independence across characteristics, which may not be valid in practical contexts. To address this constraint, augmenting Naïve Bayes by modeling attribute dependencies is suggested.

Figure 2. *Naïve Bayes Classifier*

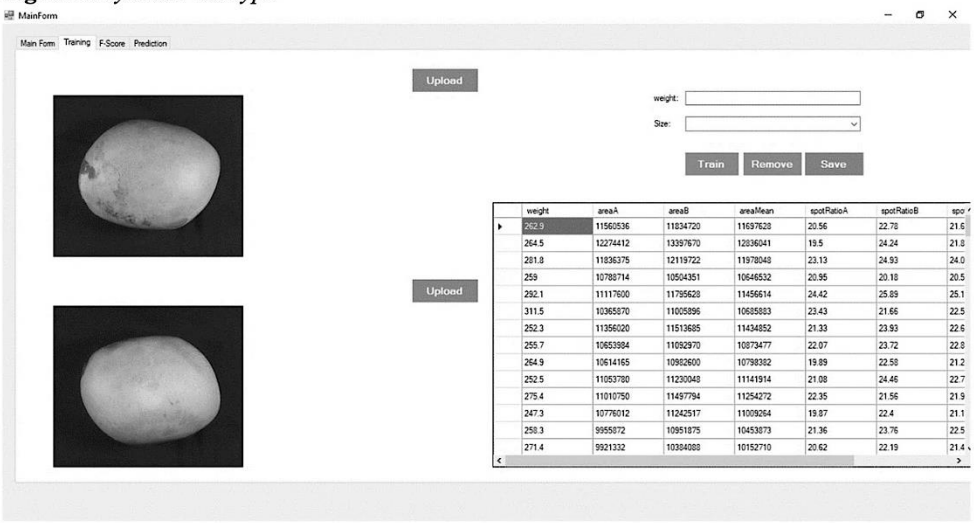


In this study, it employed the Naïve Bayes classification model to classify the carabao mango according to their grade. Blum [21] expounds on the machine learning theory, delving into the computational process underlying learning. This theory endeavors to discern the fundamental principles and information necessary for proficiently mastering various tasks, along with the essential algorithmic foundations that empower computers to learn from data and enhance performance through feedback. Its overarching goal is to refine automated learning techniques and offer insights into the fundamental mechanisms of the learning process. Using the carabao mango's size, weight, area, and spot ratio, machine learning techniques, including supervised learning, can help classify carabao mangoes based on these features. By training a model on categorized data, the mango grade can be predicted based on its physical attributes. The underlying theory informs how to design and optimize learning algorithms for practical application.

3.0 METHODS AND MATERIALS

The study utilized the quantitative-experimental research design. This design requires data collection and interpretation. It is excellent in determining trends and averages, generating predictions, assessing relationships, and generalizing conclusions. It also evaluates objective hypotheses by analyzing variable relationships [22]. Meanwhile, an experimental approach is a scientific technique for identifying the relationships between at least 2 variables. Independent and dependent variables are empirically linked to determine the relationship's nature and degree. This evaluation helps establish a cause-and-effect relationship and is also used to test hypotheses. Sirisilla [23] perceived that a scientific approach to experimental research using 2 sets of variables is achieved through this design. In this case, the first set of variables serves as a constant and is employed to calculate the second set's differences. In the research instrument, the study proposed a system for data entry and a database of carabao mangoes' records, including the fruits' size, weight, area, and spot ratio. This proposed system collected and analyzed data. Figure 3 presents the proposed system prototype.

Figure 3. *System Prototype*



Regarding the data collection, the researcher purchased 300 carabao mangoes and divided into 4 grade classes: large, medium, small, and rejected by the human grader. After classifying the carabao mangoes, a digital image of the fruit was taken using the digital single-lens reflex (DSLR) camera Nikon D3400. The image acquisition setup used an improvised chamber with a measure of 2 feet by 18 inches by 12 inches (24x18x12), composed of two 7 watts of LED lights for lighting and a digital camera at the top of the chamber. The background on all sides is black color. When the pictorial was done, mango images were transferred from digital camera memory to laptop hard disk storage for the data safekeeping and database entry purposes. The carabao mango was also weighed separately using a digital scale to determine the weight in grams according to each category. The data set has 250 images. Meanwhile, each weight and size, such as large, medium, small, or rejected carabao mangoes, are recorded using a spreadsheet application. Each mango's weight is based on grams that were weighed individually. The researcher recorded each piece of data to ensure its accuracy. Table 1 presents 150 images of carabao mangoes for the training phase and 100 for testing and classification according to their grade.

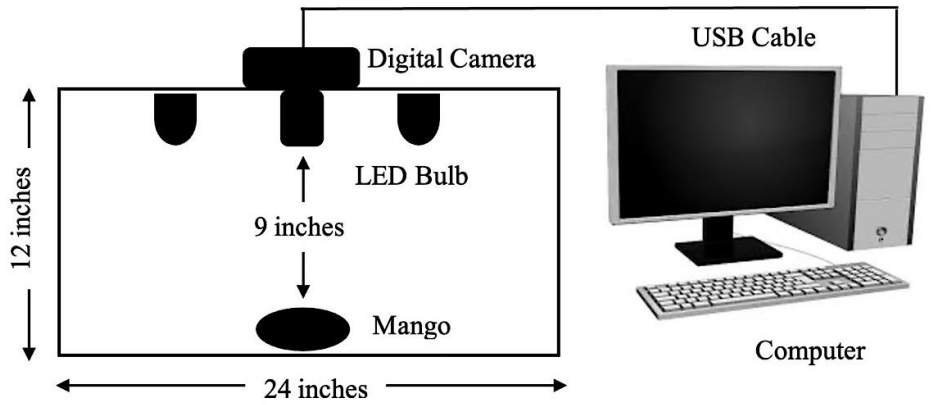
Table 1. *Training and Testing Dataset Distribution*

	Strata	Population	Total
Training Dataset	Grade A	50	50
	Grade B	50	50
	Grade C	50	50
Testing Dataset	Grade A	25	25
	Grade B	25	25
	Grade C	25	25
	Grade R	25	25

The digital scale was calibrated before weighing to ensure accuracy in measuring the carabao mangoes' actual weight. Before placing the fruit on the scale, a thorough process ensured that the scale was zeroed correctly, eliminating any extraneous weights. Throughout

the weighing process, precautions were taken to maintain a draft-free and vibration-free environment, as these factors may impact the measurements' accuracy. Careful handling of the fruits was observed to prevent any damage and minimize errors. Additionally, cautious record-keeping was implemented to maintain precise and accurate documentation for each carabao mango, ensuring a reliable dataset for future reference. The weight was recorded in grams. Figure 4 shows the layout of the image acquisition setup.

Figure 4. *Image Acquisition Setup*



The camera used in image acquisition was the Nikon D3400, an entry-level digital single-lens reflex (DSLR) camera. Table 2 presents the format for capturing the carabao mango images.

Table 2. *Digital Camera Setup*

Camera Setup	
Exposure Time [s]	1/60 Seconds
F-Number	4.5
ISO speed ratings	ISO-320
Shutter speed [s]	1/125
Aperture	4.4
Flash	No Flash
Focal length [mm]	35 mm
Color space	Srgb
Compression setting	Fine
White balance	Auto

A uniformly diffused illumination chamber was used, and the distance from the digital camera to the sample was kept constant while capturing clear carabao mango images. The capture image scene from the experimental setup has the mango image, including the background. In the segmentation process, black was used for the environment. The software is developed utilizing the C# programming language along with Aforge.NET routines. This encompasses all procedures for the acquired images, including image enhancement, noise reduction, color separation of the desired object from the background, and classification process implementation.

Modules and components should be considered when designing the system. The design phase involves strategizing the solution to the problem, transitioning from the problem to solution domain. The phase's outcome is the creation of a design document. The architectural design establishes a system's basic structure and communication between these components. The 1-tier architecture best suits the proposed method and executes in one system where everything resides in a single program.

Aforge.NET is a powerful open-source C# framework which meets the varied requirements of developers and researchers in advanced domains such as computer vision and AI. Its extensive functionalities encompass image processing, neural networks, genetic algorithms, fuzzy logic, machine learning, and robotics. By providing a flexible and accessible environment, Aforge.NET empowers users to probe into complex aspects of these domains, fostering innovation and accelerating the development of intelligent systems. A user-friendly architecture facilitates seamless integration into projects, making it a go-to resource for those seeking to push the boundaries of technology in diverse applications across vision-based computing and AI [24]. The *Aforge.Imaging* is a filter namespace serving as an integral segment of the Aforge.NET framework. Within this namespace, the developers access a comprehensive set of tools that facilitate various transformations on source images. This flexibility empowers developers to implement a wide range of image manipulations, from basic enhancements to complex filtering operations, ensuring the framework's adaptability in addressing diverse image processing needs within applications.

Aforge.Video.DirectShow namespace is also an integral to the Aforge.NET framework. This features classes designed to access video sources seamlessly through the DirectShow interface. DirectShow is a multimedia framework in Microsoft Windows that provides an abstraction layer for various multimedia components. This facilitates developers in creating C# applications with robust video capabilities by offering a convenient interface for integrating and interacting with DirectShow-compatible devices, such as webcams and capture cards. It streamlines the process of video source management, enabling efficient video capture and processing, making it an essential tool for applications requiring advanced multimedia functionalities.

Feature extraction is measuring or calculating the features from an image sample sufficient to distinguish one type of image from another. To classify the carabao mango, the characteristics of a single mango need extracting. Features are computed for subsequent examination, playing a vital role in a computer vision system by providing valuable information for image understanding, interpretation, and object categorization. During this procedure, the extracted features constitute feature vectors that undergo classification to identify the input. These feature vectors distinctly and accurately delineate the object's shape. Feature extraction endeavors to enhance the recognition accuracy by deriving relevant features. Features extracted comprise the appropriate information from the data that has an input and then the desired job with the help of using data. The extraction of components is used to decrease the source to a suitable amount from the extensive data set. Meanwhile, image processing is a significant area of feature extraction. The algorithms detect the image's desired portion [25].

After the pre-processing phase, the whole segmented mango is taken, and spot pixels are removed. The black and brown pixels inside the mango area are taken as spot pixels. The percentage of the spot pixels to the mango pixels is assumed in equation (1). The spots ratio values the categorization into less, average, and more spots for a mango. To predict the grade of carabao mango, size, weight, area, and spot ratio should be determined. The mango area can

be determined by multiplying the image's length and width or counting the number of mango pixels. The mango spot pixels were divided over the total number of pixels to determine the spot ratio.

Formula 1. *Spots Ratio*

$$Spots\ Ratio = \frac{Spots\ Pixels}{Total\ Mango\ Pixels}$$

Black spots or scratches are quality attributes used by farmers to determine the quality of the fruit [26]. Figure 5 shows the mangoes' sample images with surface defects.

Figure 5. *Images of Mango (a) Smooth Surface (b-d) With Surface Defects*

Euclidean distance was used for color filtering to determine the area and spot ratio. Below is the formula for getting the length specified in Formula 2.

Formula 2. *Euclidean Distance Equation*

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + (p_3 - q_3)^2}.$$

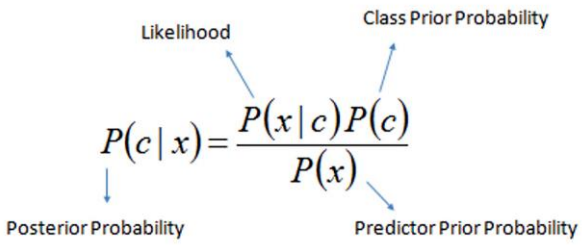
Getting the distance between two colors multiplies each channel's difference between them and then adds it all together. Identifying the distance between 2 colors is trivial when searching a set array of colors and looking for the nearest match. The closest match is the color with the lowest distance. A difference of zero means that the colors are a perfect match.

The Naïve Bayes classified the carabao mangoes using the fruits' size, weight, area, and spot ratio. Using the features extraction method, the input image of fruits was analyzed. Naïve Bayes operates as a probabilistic classifier grounded on Bayes' Theorem, incorporating the Naïve (Strong) independence assumption. These classifiers assume that the impact of a variable's value on a particular class remains independent of the values of other variables. This assumption is termed as class conditional independence.

Bayes' theorem provides a means to calculate the posterior probability, P(c | x), using P(c), P(x), and P(x | c). The Naïve Bayes classifier operates under the assumption that the effect

of the predictor value (x) on a particular class naïve is independent of the values of other predictors. This assumption is known as class conditional independence [27].

Formula 3. *Naïve Bayes Equation*



$$P(c \mid X) = P(x_1 \mid c) \times P(x_2 \mid c) \times \cdots \times P(x_n \mid c) \times P(c)$$

Where:

P (c | x)	= is the posterior probability of class (target) given predictor (attribute)
P (c)	= is the prior probability of class
P (x | c)	= is the likelihood, which is the probability of predictor given class
P (x)	= is the prior probability of the predictor

In plain English, the above calculation can be written as:

$$Posterior = \frac{Prior\ x\ Likelihood}{Evidence}$$

Moreover, data analysis was employed to classify 250 carabao mango images. After the preparation and image were ready, the photos underwent pre-processing and extraction, followed by classification. The variables' collected values, size, weight, area, and spot ratio were organized and prepared for data analysis. Due to the increasing demand for this commodity in both local and international markets, the Philippine National Standard for Fresh Fruits – Mangoes, PNS/BAFPS 13:2004, was initially established in July 2001. This standard was developed under the Technical Assistance on Safety and Quality Standards Covering Products of High-Value Commercial Crops by the Bureau of Agriculture and Fisheries Product Standards (BAFPS) [28]. Table 3 shows the size classification of carabao mango fruits.

Table 3. *Size Classification of Carabao Mango*

Size Classification	
Size	**Weight (g)**
Large	300 - 349
Medium	250 - 299
Small	200 - 249

Technium Vol. 22, pp.14-32 (2024)
ISSN: 2668-778X
www.techniumscience.com

The schematic diagram of problem one (Figure 6) uses the input of the size, weight, area, and carabao mango spot ratio. The Naïve Bayes classifier classified grades A, B, C, and R. The result of the classification model determined the accuracy of the Naïve Bayes classifier proposed system.

Figure 6. *Determining the Grade of Carabao Mangoes*

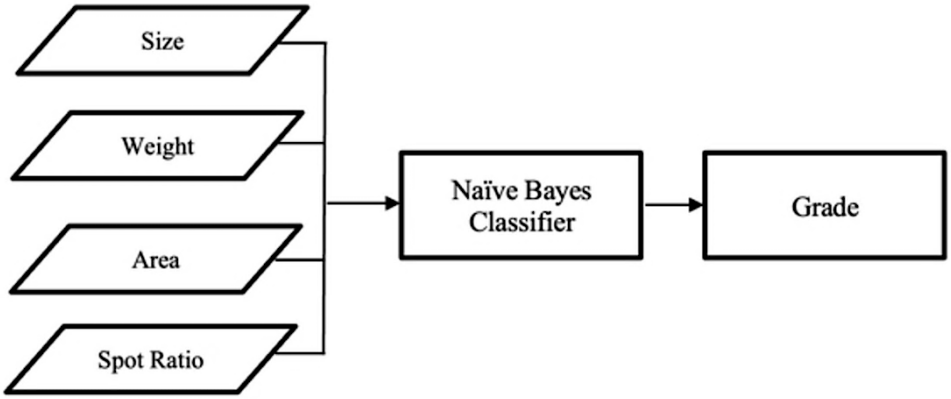
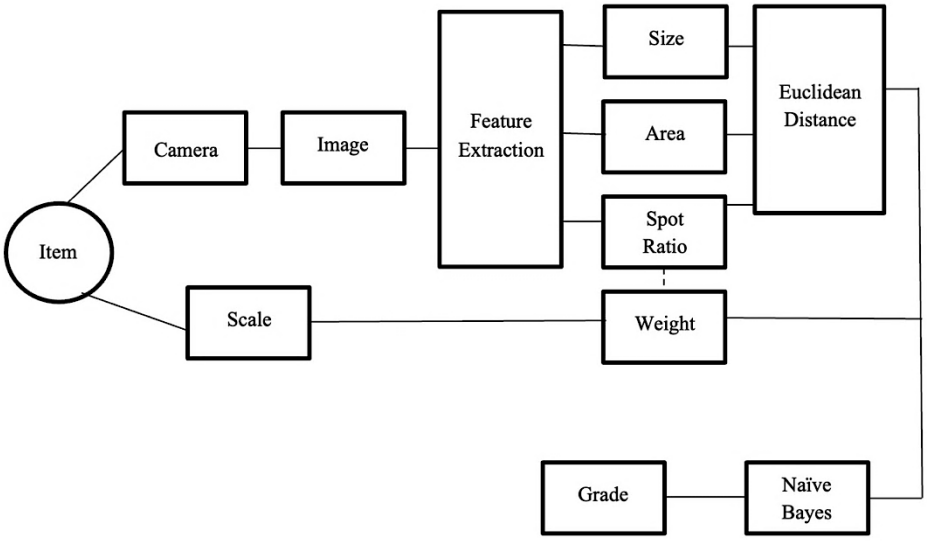


Figure 7 illustrates the schematic diagram that shows the method for classifications of carabao mango by applying the variables size, weight, area, and spot ratio of fruits. The input for the model is the captured images of mango fruits. The features extraction model was applied: Euclidean distance, and for the classification, the Naïve Bayes classifier was used for the category result according to grades like grade A, grade B, grade C, and grade R.

Figure 7. *Feature Extraction*



The schematic diagram in Figure 7 shows that weight, size, and area are the basis for determining grades A, B, and C. For a grade, R is the spot ratio of the carabao mango. To verify the consistency of the Naïve Bayes classifier in grading the carabao mango fruits using the size, weight, area, and spot ratio, the following equation (Formula 4) was used:

Formula 4. *Accuracy Formula*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} 100\%$$

Classification precision is frequently employed as it offers a concise summary of the model's effectiveness. The F-Measure provides a means to consolidate and encapsulate all accuracies into a single metric that encompasses all aspects. After precision and recall have been computed for a binary or multiclass classification problem, these 2 metrics can be amalgamated into the F-Measure calculation [29]. The F-Measure, also known as the F-Score, gauges a test's accuracy, representing the harmonic mean of the test's precision and recall.

Formula 5. *F-Measure Formula*

$$F_1 = \left(\frac{\text{recall}^{-1} + \text{precisior}^{-1}}{2} \right)^{-1} = 2 \cdot \frac{\text{precisior} \cdot \text{recall}}{\text{precision} + \text{recall}}.$$

With the accuracy and recall of a test considered, the positive outcomes are precision, often referred to as the positive predictive value. Memory, also known as sensitivity, reflects a test's capacity to correctly identify positive outcomes, thereby achieving the correct positive rate. The F-Score attains its optimum value of 1, indicating flawless accuracy and recall. Conversely, a score of 0 represents the poorest F-Score, signifying the lowest precision and recall. The F-Score serves as a metric for evaluating a test's accuracy. It balances and remembers the use of accuracy to do so. Using precision and recall, the F score can provide a

Formula 6. *Recall Formula*

$$Recall = \frac{TP}{TP + FN}$$

Formula 7. *Precision Formula*

$$Precision = \frac{TP}{TP + FP}$$

more objective measure of a test's success. The F-Score is also used for information retrieval to calculate search, document classification, and query classification. The total figure of positive examples correctly classified to the total number of positive models can be defined as a recall. High recall suggests the proper recognition of the class (a small number of FN). In terms of precision, divide the number of positive cases correctly classified by the number of positive models expected for accuracy. High precision shows that a bar marked as positive is positive (a small number of FP).

In instance like high recall, low precision, this suggests that most positive examples are understood correctly (low FN), but some false positives exist. In low recall, high precision, it illustrates the lack of positive examples (high FN), but according to Nitin [36], those foreseen as positive are indeed positive (low FP). Regarding descriptive analysis, 250 carabao mangoes were classified according to grade: grade A (large), grade B (medium), grade C (small), and grade R (reject). Each class contains 75 carabao mangoes divided into testing and training, 50 for each training for class A, class B, and class C, while 25 for testing each class for class A, class B, class C and class R. The precision, recall, and f-measures percentage was computed to identify the accuracy of the classifier.

4.0 RESULTS AND DISCUSSION

Training Data Set

The data set, comprised of carabao mangoes, was accurately curated and utilized as the initial training material for the proposed system. This collection, specifically chosen for its relevance and variety, enhanced the system's ability to recognize, analyze accurately, and process characteristics unique to carabao mangoes. Table 4 shows that the overall accuracy of the categorization model was 95%, which is excellent. The model correctly identified 95% of the mangoes in the dataset. This positive result indicates that the model can be used to classify carabao mangoes accurately. In the training data set, 150 carabao mangoes were categorized, and 143 were successfully classified, accounting for about 95% of the overall data set. Only 7 were wrongly classified, accounting for 5% of the available data set. Supekar and Wakode [30] developed a computer vision-based mango grading system, achieving 99.20% accuracy in size classification and 100% accuracy in ripeness and shape-based categorization, demonstrating the potential of computer vision technology in improving fruit industry efficiency and consistency.

Table 4. *Summary of Instances for Training Data Set*

Classified Instances	Training Data Set	Percentage
Correctly Classified Instances	143	0.95
Incorrectly Classified Instances	7	0.05

Detailed Accuracy by Class in terms of Precision, Recall, and F-Measure

Table 5 shows the precision, recall, F-Measure, and class results for each group of data sets. The precision for class A carabao mango, in this case, is 94%, which is considered good precision. The precision for class B carabao mango is 96%, which is still regarded as adequate, and class C carabao mango precision also got a 96% score. A classification model's precision measures how accurate the model is at predicting the correct class for a given data point. A

high precision indicates that the model is effective at identifying the suitable class, whereas a low precision suggests that the model is ineffective. Overall, the classification model for carabao mangoes is accurate. For the most part, the model can correctly identify the class of mangoes.

In recall, class A had an accuracy of 96%, class B scored 91%, and class C had the highest result, which was 100%. This is a good result because it shows that the classifier can correctly identify carabao mangoes of various classes. This can be helpful in multiple applications, including quality control, sorting, and marketing. Particularly, recall measures a classifier's ability to correctly identify all instances of a specific class. Carabao mangoes are classified into 3 categories: class A, class B, and class C. The recall for each type indicates that the classifier accurately identified all occurrences of that class with high accuracy. The recall for each carabao mango class is positive, indicating that the classifier correctly identified most mangoes in each category. Class C had the best recall, followed by classes A and B.

Relative to F-measure, in carabao mango classification task, the results indicate that class C has the highest F-measure score of 98%, followed by class A with a score of 95% and class B with a score of 93%. This suggests that class C is the most accessible class to classify, followed by class A and then class B. The F-measure results indicate that the classifier can identify carabao mangoes accurately. The F-measure measures a model's accuracy on a dataset and is used to evaluate binary sorting systems, which classify instances into 'positive' or 'negative.' The F-measure is a way of merging the model's precision and recall, which is well-defined as the vocal mean of the model's precision and recall. The F-measure results also suggest that the classifier is more accurate at classifying some classes of mangoes than others.

The paper by Kapila et al. [31] proposes a support vector machine (SVM) classifier utilizing hierarchical features taken from the entirely associated layer of the pre-trained deep convolutional neural network as a classification model for various apple fruit kinds. The accuracy, precision, recall, and F1-score of the suggested classification model are evaluated in performance metrics. The ResNet 50 Pre-trained deep neural network's features were used to train the SVM classifier, which achieved 99.1% accuracy and precision, 95.4% F1-score, and 98.6% recall.

Table 5. *Detailed Accuracy by Class*

Precision	Recall	F-Measure	Class
94%	0.96	0.95	Class A
96%	0.91	0.93	Class B
96%	1.00	0.98	Class C

Confusion Matrix for Training Data Set

Table 6 presents the findings for class A, which show that 47 out of 50 data sets are correctly categorized, while 3 are wrongly classified. This implies that the model has a high accuracy rate for class A, successfully detecting most carabao mangoes. The confusion matrix for class B reveals that 48 out of 50 data sets are correctly identified, indicating a reasonable accuracy rate. This implies that the model fared well in classifying carabao mangoes, with only 2 instances of misclassification. This shows that the model distinguishes class B carabao mangoes from other classes. For class C, the findings show that 48 out of 50 data sets are correctly categorized. This indicates that the model's accuracy for class C is reasonable. This is a promising result of the model being used in classifying carabao mangoes.

Meanwhile, the confusion matrix is a tabular representation utilized to assess the effectiveness of a classification model. It illustrates the count of instances that were accurately classified as well as those that were erroneously classified. In support, Abidin et al. [32] gives a tabular confusion matrix describing the model's performance on previously tested data. By studying these indicators within the confusion matrix, insights can be gained in the categorization model's strengths and flaws. These insights help us fine-tune our algorithms, improve accuracy, and make educated decisions based on the model's performance.

Table 6. *Confusion Matrix for Training Data Set*

A	B	C	Actual Classification
47	3	0	Class A
2	48	0	Class B
0	2	48	Class C

Instances for Testing Data Set

This data collection was specifically distributed during the proposed system's testing phase. It provided a vital benchmark for evaluating the system's performance under real-world situations, guaranteeing its accuracy, dependability, and efficacy in handling carabao mango in various scenarios. Table 7 describes a classification algorithm's performance on a test dataset, including its accuracy in classifying cases. It categorizes the results into "Correctly Classified Instances" and "Incorrectly Classified Instances." Here, the algorithm correctly ranked 93 of the 100 cases in the testing set, resulting in a 93% accuracy rate. The incorrectly classified instances count is 7, corresponding to a 7% error rate. This error rate, while relatively low, indicates that there is still an area for improvement in the model's predictive capabilities.

This high percentage of correct classifications implies that the algorithm is extremely good at identifying and categorizing data as intended, indicating strong model performance in this context. The table gives valuable information about the model's capabilities and potential areas for improvement in accuracy and reliability in categorizing instances within the dataset. Win [33] suggests that the ideal classification approach should exhibit high accuracy, calculated as the ratio of correctly identified fruit samples to the total number of samples employed during testing. The system's effectiveness is assessed by employing five distinct fruit varieties not present in the dataset. These samples are then subjected to testing using the system, which evaluates 20 photographs from each mango type. The database is trained using 50 images from each mango fruit variation. According to the study, Naïve Bayes offers good promise for detecting mango types in a nondestructive and accurate manner.

Table 7. *Summary of Instances for Testing Data Set*

Classified Instances	Testing Data Set	Percentage
Correctly Classified Instances	93	0.93
Incorrectly Classified Instances	7	0.07

Confusion Matrix for Testing Data Set

Table 8, the given confusion matrix, represents the performance of a classification model across four classes: A, B, C, and R. Each row shows the actual class. In contrast, each column represents the predicted class. For class A, all 25 instances are correctly classified, indicating perfect classification with no false positives or negatives. Similarly, class B has 25 instances correctly classified without any misclassification. Class C, however, shows some misclassification. Of 25 cases, 18 are correctly classified as class C, but seven are incorrectly classified as class B. This indicates confusion between classes B and C, reducing the model's precision and recall. Class R is ideally classified with all. 25 instances were correctly identified, showing no misclassification.

The model demonstrates high accuracy for classes A, B, and R with perfect classification. However, the performance on class C could be more accurate due to confusion with class B, indicating a potential area for improvement in the model's ability to distinguish between these two classes. The model's overall performance is strong but could benefit from improvements in determining class C from class B. In Wenzhong [34], the deep learning network named IntelFruit was trained using a fruit dataset. Subsequently, the model underwent evaluation on the test set and demonstrated satisfactory performance. The outcomes of classification were illustrated through a confusion matrix, wherein each row corresponds to the actual category, and each column signifies the predicted outcome. The findings were used to visually evaluate the classifier's performance and identify the highlighted classes and network model features.

Table 8. *Confusion Matrix for Testing Data Set*

A	B	C	R	Actual Classification
25	0	0	0	Class A
0	25	0	0	Class B
0	7	18	0	Class C
0	0	0	25	Class R

Depth of Analysis

The analysis presents an insightful examination of the performance of the Naïve Bayes classification model used for categorizing carabao mangoes based on a training data set and subsequent testing with a different data set. This assessment evaluates the model's accuracy, precision, recall, F-measure, and confusion matrix results, collectively providing a comprehensive overview of the model's effectiveness and areas for enhancement.

In training data set analysis, specifically in *accuracy,* the training data set comprised 150 carabao mangoes, with 143 correctly classified, yielding an accuracy rate of 95%. This high accuracy indicates the model's robustness in identifying carabao mangoes, suggesting that it has learned the distinguishing features of the mangoes effectively during the training phase. In terms of the *precision and recall,* the model demonstrated high accuracy for all classes (A: 94%, B: 96%, C: 96%), indicating its ability to minimize false positives effectively. The recall rates (A: 96%, B: 91%, C: 100%) suggest that the model can identify most true positives, particularly for class C, which achieved perfect recall. These metrics highlight the model's competency in class differentiation and its potential utility in applications like quality control and sorting. Relative to the *F-Measure,* the results (A: 95%, B: 93%, C: 98%) reflect the model's balanced performance between precision and recall, with class C being the easiest to classify accurately. This balance is crucial for maintaining overall model effectiveness across various classes. Meanwhile, the *confusion matrix* for the training data set reveals high accuracy

rates for all classes, with minimal misclassification. This suggests that the model is generally effective at distinguishing between the classes of carabao mangoes.

In training data set analysis specifically in *accuracy*, the testing data set was slightly lower at 93%, with 93 out of 100 mangoes correctly classified. This minor drop indicates new variations or unseen instances in the testing set that needed to be fully represented in the training data, highlighting the diverse and comprehensive training set. Regarding the *misclassification analysis*, the 7% error rate in the testing data set suggests areas for improvement. Misclassifications could stem from data noise, inadequate features, or the model's limitations in capturing complex patterns, emphasizing continuous model refinement. In *confusion matrix*, the testing set reveals perfect classification for classes A, B, and R but shows some confusion between classes B and C, with seven instances of class C being misclassified as class B. This indicates a specific challenge in differentiating these 2 classes, pointing to a potential area for model improvement.

This study thoroughly evaluated the Naïve Bayes classification model's performance in categorizing carabao mangoes into class A, class B, and class C. While the training data set achieved an impressive 95% accuracy, a modest reduction to 93% in the testing data emphasizes the significance of diverse training data to manage unknown variables. Despite great precision and recall in most classes, the confusion matrix in the testing data highlighted distinct flaws. The model struggled to distinguish between classes B and C, indicating the need for additional refining. The model has good promise for carabao mango categorization, but further enhancements are needed to overcome misclassification concerns and improve generalizability. The basis for classification is the carabao mango's size, weight, area, and spot ratio.

According to Miriti [35], Naïve Bayes has a good chance of accurately and nondestructively identifying apple varietals. Even though this system cannot equal the precision and accuracy of the human eye and hand, it can undoubtedly outperform them in speed and cost. The study examined how well the Naïve Bayes classifier performed compared to other methods previously employed to address the apple classification and recognition challenge. The comparison criteria were the extracted characteristics, the classifiers used, and the accuracy attained.
When comparing their classification accuracy with that of the Naïve Bayes technique, it was found that Naïve Bayes achieved a higher accuracy rate of 91%, whereas principal components analysis, fuzzy logic, and multi-layer perception achieved accuracies of 90%, 89%, and 83%, respectively.

5.0 CONCLUSION

The result showed that the study classifies the carabao mango depending on the class. The carabao mango classifier utility model extracted the mango's feature through image processing, and each mango is classified using the Naïve Bayes classifier. The Naïve Bayes classifier detected the class of mangoes and classified it into three classes, namely: A (large), B (medium), and C (small). The accuracy of the utility model was 95%, indicating its effectiveness in correctly classifying mango classes. A confusion matrix table was used to evaluate the performance of the classification utility model.

6.0 LIMITATION OF THE FINDINGS

The Naïve Bayes model for carabao mango classification showed promise in classifying the mango grade. However, limitations exist. The training dataset (150 mangoes) might limit generalizability, and features were restricted to size and weight. Future work should address these limitations by expanding the dataset size and incorporating additional features like shape, color, and ripeness. Testing in diverse conditions and comparing with other models are also crucial for further development and maintaining the model's effectiveness.

7.0 PRACTICAL VALUE OF THE PAPER

The Naïve Bayes model effectively classified carabao mangoes, achieving 95% accuracy. This technology offers practical applications in agriculture and food. Integration into automated sorting and grading systems can streamline processing, reduce costs, and ensure consistent quality for consumers. Additionally, data collected by the model can provide insights into consumer preferences and supply chain optimization. Analysis of mango characteristics can further inform research on cultivation practices and new mango varieties.

8.0 DIRECTIONS OF THE FUTURE RESEARCH

The Naïve Bayes utility model accurately determined the carabao mangoes by size, weight, area, and spot ratio features. However, limitations exist, including misclassification rates and sensitivity to mango variations. Future research should explore deep learning for advanced feature extraction and expand the dataset to include other mango varieties. Deploying the model on farm IoT devices could enable real-time monitoring and selective harvesting. Open-sourcing the model could benefit small-scale farmers by democratizing AI technology within the agricultural and food industry.

9.0 ACKNOWLEDGEMENT

The main author acknowledges the support of Dr. Violeta V. Guillergan and Katrina Charliz V. Guillergan in accomplishing this paper. He also thanks Eleazar Comprendio and Genelove G. Comprendio, Green Apple P. Guillergan and Derin Abubakir M Amin Mohammed Amin for the financial support. Prof. Filomeno S. Posadas Jr. for constructing the image acquisition setup, Jelry P. Fuentes, and Arden D. Nobleza for their assistance in preparing the data sets, and all the people who contributed their expertise and assistance in accomplishing this research.

10.0 REFERENCES

[1] Siano AB, Aya RAM. Technology to identify genuine 'carabao' mango develop by VSU. https://www.pcaarrd.dost.gov.ph/index.php/quick-information-dispatch-qid-articles/technology-to-identify-genuine-carabao-mango-developed-by-vsu

[2] Ongkiko R. UPLB takes part in the advancement of the Philippine mango industry. 2015. https://www.ovcre.uplb.edu.ph/press/features/item/427-uplb-takes-part-in-the-advancement-of-the-philippine-mango-industry

[3] Gonzalez G. What are the sweetest mangoes. 2022. https://sweetishhill.com/what-are-the-sweetest-mangoes/

[4] Ardepolla JA, Cortez MJ, Escorpion AL, Adtoon JJ. Identification and classification of export quality carabao mangoes using image processing. InProceedings of the 6th

International Conference on Bioinformatics Research and Applications 2019 Dec 19 (pp. 13-17). https://doi.org/10.1145/3383783.3383785
[5] Tababa J. The Future of Mango Cultivation: Advancements and innovations. 2023. https://mb.com.ph/2023/7/4/the-future-of-mango-cultivation-advancements-and-innovations
[6] Liu J, Kong X, Xia F, Bai X, Wang L, Qing Q, Lee I. Artificial intelligence in the 21st century. Ieee Access. 2018 Mar 26;6:34403-21. https://doi.org/10.1109/ACCESS.2018.2819688
[7] Ramasubramanian K, Singh A, Ramasubramanian K, Singh A. Machine learning theory and practices. Machine Learning Using R. 2017:219-424. https://doi.org/10.1007/978-1-4842-4215-5_6
[8] Swamynathan M. Mastering machine learning with python in six steps: A practical implementation guide to predictive data analytics using python. Manohar Swamynathan; 2017. http://aisel.aisnet.org/sjis%0Ahttp://aisel.aisnet.org/sjis/vol19/iss2/4
[9] Sammut C, Webb GI. Encyclopedia of machine learning and data mining. Springer Publishing Company, Incorporated; 2017 Mar 15.
[10] Taboga M. Bayesian Inference: Lectures on probability theory and mathematical statistics. Kindle Direct Publishing. 2021. https://www.statlect.com/fundamentals-of-statistics/Bayesian-inference
[11] Kaur N, Kaur R. Content Based Image Retrieval Using Color Mean with Feature Classification Using Naïve Bayes. International Journal of Advanced Research in Computer Science. 2016 Nov 1;7(6).
[12] Oliver J. Philippine mango industry roadmap 2017-2020. Journal of Chemical Information and Modeling, 53(9), 1689–1699. 2018.
[13] Department of Agriculture (DA). Philippine mango industry roadmap 2017-2022. 2022. https://www.da.gov.ph/wp-content/uploads/2019/06/Philippine-Mango-Industry-Roadmap-2017-2022.pdf
[14] Pauly L, Sankar D. A new method for sorting and grading of mangos based on computer vision system. In2015 IEEE International Advance Computing Conference (IACC) 2015 Jun 12 (pp. 1191-1195). IEEE. https://doi.org/10.1109/IADCC.2015.7154891
[15] Chaudhari D, Waghmare S. Machine vision based fruit classification and grading—a review. InICCCE 2021: Proceedings of the 4th International Conference on Communications and Cyber Physical Engineering 2022 May 16 (pp. 775-781). Singapore: Springer Nature Singapore. https://link.springer.com/chapter/10.1007/978-981-16-7985-8_81
[16] Gill HS, Khehra BS. Fruit image classification using deep learning. Computers, Materials and Continua, 71(2), 5135–5150. 2022. https://doi.org/10.32604/cmc.2022.022809
[17] Han J, Kamber M, Pei J. Data mining concepts and techniques third edition. University of Illinois at Urbana-Champaign Micheline Kamber Jian Pei Simon Fraser University. 2012.
[18] Müller AC, Guido S. Introduction to machine learning with Python: a guide for data scientists. " O'Reilly Media, Inc."; 2016 Sep 26. https://doi.org/10.1007/978-3-030-36826-5_10
[19] Navlani A. Naive Bayes classification tutorial using Scikit-learn. 2018. https://www.datacamp.com/tutorial/naive-bayes-scikit-learn
[20] Kaviani P, Dhotre S. Short survey on naive bayes algorithm. International Journal of Advance Engineering and Research Development. 2017 Nov 11;4(11):607-11

[21] Blum A. Machine learning theory. Carnegie Melon Universit, School of Computer Science. 2007;26. http://www.cs.cmu.edu/afs/cs/user/avrim/www/Talks/mlt.pdf
[22] Kharbach M. What is quantitative research?. 2023. https://www.selectedreads.com/what-is-quantitative-research-according-to-creswell/
[23] Sirisilla S. Experimental research design. 2023. https://www.enago.com/academy/experimental-research-design/
[24] Aforge.NET. Aforge.NET framework. 2024. https://www.aforgenet.com/framework/

[25] Agilandeeswari L, Prabukumar M, Goel S. Automatic grading system for mangoes using multiclass SVM classifier. International Journal of Pure and Applied Mathematics. 2017;116(23):515-23.
[26] Nandi CS, Tudu B, Koley C. Machine vision based automatic fruit grading system using fuzzy algorithm. InProceedings of the 2014 International Conference on Control, Instrumentation, Energy and Communication (CIEC) 2014 Jan 31 (pp. 26-30). IEEE. https://doi.org/10.1109/CIEC.2014.6959043
[27] Saed S. An introduction to data science. (2022). http://www.saedsayad.com/naive_bayesian.htm
[28] Coates LT, Cooke D, Persley B, Beattie N, & Ridgway R. Postharvest diseases of horticultural product. Tropical Fruit, 2. 1995.
[29] Brownlee J. How to calculate precision, recall, and f-measure for imbalanced classification. 2023. https://machinelearningmastery.com/precision-recall-and-f-measure-for-imbalanced-classification/
[30] Supekar AD, Wakode M. Multi-parameter based mango grading using image processing and machine learning techniques. INFOCOMP Journal of Computer Science. 2020 Dec 8;19(2):175-87. http://177.105.60.18/index.php/infocomp/article/view/756
[31] Kapila G, Vandana B, Khaitan A, Francis Avinash A, Ajay Kumar CH. Apple fruit classification and damage detection using pre-trained deep neural network as feature extractor. InInnovations in Electronics and Communication Engineering: Proceedings of the 9th ICIECE 2021 2022 Mar 13 (pp. 235-243). Singapore: Springer Singapore. https://doi.org/10.1007/978-981-16-8512-5_26
[32] Abidin AA, Hamzah H, Endah M. Efficient Fruits Classification Using Convolutional Neural Network. International Journal of Informatics and Computation. 2021 Oct 29;3(1):1-9. https://doi.org/10.35842/ijicom.v3i1.31
[33] Win O. Classification of mango fruit varieties using naive Bayes algorithm. I International Journal of Trend in Scientific Research and Development (IJTSRD). 2019;3(5):1475-8. https://doi.org/https://doi.org/10.31142/ijtsrd2667
[34] Wenzhong L. Interfruit : Deep learning network for classifying fruit images. 2020. https://www.biorxiv.org/content/10.1101/2020.02.09.941039v2.abstract?%3Fcollection
[35] Miriti E. Classification of selected apple fruit varieties using Naive Bayes (Doctoral dissertation, University of Nairobi). 2016. http://hdl.handle.net/11295/97285
[36] Nitin P. Confusion matrix in machine learning. 2023. https://www.geeksforgeeks.org/confusion-matrix-machine-learning/