Jurnal Mandiri IT
Vol. 13 No. 1, July (2024), pp. 18-27

ISSN 2301-8984 (Print), 2809-1884 (Online)
Published by Institute of Computer Science

Comparison of naïve bayes and KNN for herbal leaf classification

Bangkit Indarmawan Nugroho[1], Muhammad Wazid Khusni[2], Pingky Septiana Ananda[3], Gunawan Gunawan[4]
[1,3]Information System, STMIK YMI TEGAL, Indonesia
[2,4]Informatics Engineering, STMIK YMI TEGAL, Indonesia

A R T I C L E I N F O

Article history:

Received May 30, 2024
Revised May 31, 2024
Accepted Jun 10, 2024

Keywords:

Classification;
GLCM;
Herbal Leaf;
K-Nearest Neighbor;
Naïve Bayes.

ABSTRACT

This study aims to compare the effectiveness of two classification algorithms, namely Naïve Bayes Classifier and K-Nearest Neighbor (KNN), in classifying herbal leaves. This research design uses a quantitative approach with experimental analysis and model validation. The dataset consisted of images of papaya leaves, pandanus, cat's whiskers, and betel nut taken in different lighting conditions. The methodology includes pre-processing of data by converting images into grayscale, feature extraction using Gray Level Co-occurrence Matrix (GLCM), and application of Naïve Bayes and KNN algorithms. The main results showed that KNN achieved 90.00% accuracy with precision, recall, and F1-score of 88.33% respectively, higher than Naïve Bayes which had 82.50% accuracy, 81.46% precision, 85.83% recall, and 82.27% F1-score. In conclusion, KNN is superior in the classification of herbal leaves to Naïve Bayes, although it requires a longer computational time. Further research is recommended to optimize algorithm parameters and explore the integration of deep learning techniques to improve classification accuracy and efficiency.

Corresponding Author:

Muhammad Wazid Khusni,
Informatics Engineering,
STMIK YMI TEGAL,
#1 Pendidikan Street, Tegal City, Central Java 52142, Indonesia.
Email: khusniwazid@gmail.com

1. INTRODUCTION

The utilization of herbs for medicine and health supplements has increased significantly worldwide in the last ten years (Tangkiatkumjai et al., 2020). Accurate and rapid identification of various species of herbs is becoming important, not only for medical purposes but also for biodiversity conservation (Howes et al., 2020). However, challenges arise in this identification process due to similarities between species and varieties that can be confusing even to botanists (Upton et al., 2020). In this context, the development of an efficient classification algorithm to identify herbal leaves based on their morphological features becomes very important (Sachar & Kumar, 2021).

The main problem in the classification of herbal leaves is the high dimensionality of the data and the similarity of traits between different species, which can result in a decrease in classification accuracy (Naeem et al., 2021). This research is important because proper classification algorithms can help in the faster and more accurate introduction of herbal species, further supporting the preservation, research, and sustainable use of herbs (Meng et al., 2023).

Previous research has shown the great potential of machine learning and image processing techniques in identifying and classifying medicinal plants and fruits (Mulugeta et al., 2024). V. Kamath uses algorithms such as SVM, Naive Bayes, KNN, and Decision Trees to categorize Ayurvedic herbs with significant results (Kamath, 2024). K. Pushpanathan et al. showed that SVM, Naive Bayes, KNN, and CNN can improve the accuracy of identification of medicinal

plants (Pushpanathan et al., 2021). S. Sachar and A. Kumar highlight the importance of feature extraction and classification techniques using leaf images with similar algorithms (Sachar & Kumar, 2021). S. K. Behera et al. used machine learning and transfer learning to classify papaya fruit ripeness, with the VGG19 model achieving 100% accuracy (Behera et al., 2021). F. Twum et al. used the Log Gabor filter for texture analysis in the identification of medicinal plants, showing accuracy of up to 98% (Twum et al., 2022).

Previous research has shown the effectiveness of various machine learning algorithms in classifying medicinal plants, often limited by challenges such as various lighting conditions and high data dimensions. This study addresses this gap by utilizing the Gray Level Co-occurrence Matrix (GLCM) for feature extraction and comparing the performance of the Naïve Bayes Classifier and K-Nearest Neighbor (KNN) algorithms. By focusing on this method, the study aims to improve classification accuracy and efficiency under different lighting conditions, providing a more robust solution for herbal leaf identification. A comprehensive comparison between these two algorithms is expected to offer valuable insights into the optimal approach to classifying herbal leaves.

This study aims to compare the effectiveness of two popular classification algorithms, namely Naïve Bayes Classifier and K-Nearest Neighbor (KNN), in classifying herbal leaves. The Naïve Bayes Classifier and K-Nearest Neighbor (KNN) methods were chosen because of their respective advantages in various classification scenarios and their ability to handle big data. Naïve Bayes is advantageous because of its simplicity, efficiency, and effectiveness in handling high-dimensional data where feature independence is assumed. KNN, known for its simplicity and effectiveness in non-linear data distribution, excels in scenarios where data distributions are complex and do not conform to parametric assumptions, making them suitable for diverse data sets.

This study specifically identifies and classifies specific herbal species rather than classifying herbal leaves in general. The dataset included images of papaya leaves, pandan leaves, cat's whisker leaves, and betel leaves, and the goal was to compare the performance of the Naïve Bayes and K-Nearest Neighbor algorithms in classifying specific types of herbal leaves based on their morphological features. By focusing on this specific species, the study intends to determine the most suitable algorithm for accurate classification in a given context, contributing to the development of a proper automatic identification system for these herb leaves. This targeted approach ensures that these findings are directly applicable to real-world scenarios involving this particular ingredient.

Addressing this issue will provide important insights into the application of classification algorithms in botany and may aid in the development of a more accurate automatic identification system for herbal leaves (Abraham & Kellogg, 2021). This study is expected to fill in the gaps in the literature by providing a comprehensive comparison between two different classification techniques in this particular context (Edison et al., 2021).

The high dimensions and similarities between species affect the accuracy of classification because they introduce noise and redundancy, making it difficult to distinguish between the same species and often leading to misclassification. The practical application of this research lies in the development of reliable and efficient systems for the identification of herbal species, which are essential for drug use, conservation efforts, and agricultural management. The results will guide the selection of the right algorithm for the classification of herbal leaves, improving the accuracy and speed of the identification process in real-world applications.

Some researchers focus on image classification techniques for herbal leaves and leaf diseases, but are limited to the integration of methods in different lighting conditions (Kolhar & Jagtap, 2023). Therefore, this study will apply the extraction of the Gray Level Co-occurrence Matrix (GLCM) feature and the application of the Naïve Bayes Classifier and K-Nearest Neighbor (KNN) algorithms used to classify several types of herbal leaves (Wu et al., 2023), namely papaya leaves, pandan leaves, cat's whisker leaves, and betel leaves. The purpose of this study is to compare the two methods to see better performance and evaluate performance to determine the most appropriate algorithm in classifying the image of herbal leaf types.

2. RESEARCH METHOD

This research has several stages, where two classification algorithms, Naïve Bayes Classifier and K-Nearest Neighbor (KNN) will be applied to the same dataset. Here are the research steps.

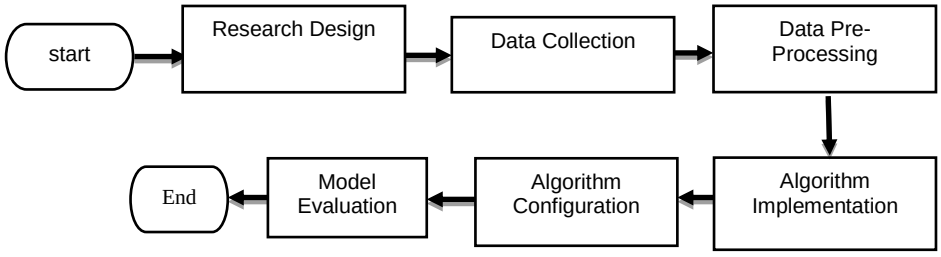


Figure 1. Research flow

Research Design

This research uses a quantitative approach with a combination design of experimental methods, quantitative analysis, and model validation (Falk et al., 2023). This design was chosen to provide a comprehensive understanding of the performance of each algorithm in the classification of herbal leaves (Naeem et al., 2021).

Data Collection

The image data of herbal leaves used are papaya leaves, pandan leaves, cat's whisker leaves, and betel leaves. The data was taken from the results of portraits using mobile phones with a total of 200 images of herbal leaves where each type of leaf consists of 50 images.

Data Pre-processing

The pre-processing or data processing stage, the first step is to reduce the pixels in the leaf image from 2992 x 4000 pixels to 612 x 818 pixels, then the image image is converted into grayscale. Once the data is converted into grayscale, calculate feature extraction using the Gray Level Co-occurrence Matrix (GLCM) and split the image data into separate directories. The first directory is the training data used in training the classification model, while the second directory is the test data used to evaluate the performance of the model.

Algorithm Implementation

The Naïve Bayes Classifer and K-Nearest Neighbor (KNN) algorithms will be implemented using the pyhton programming language with pre-processed data (Ullah et al., 2021).

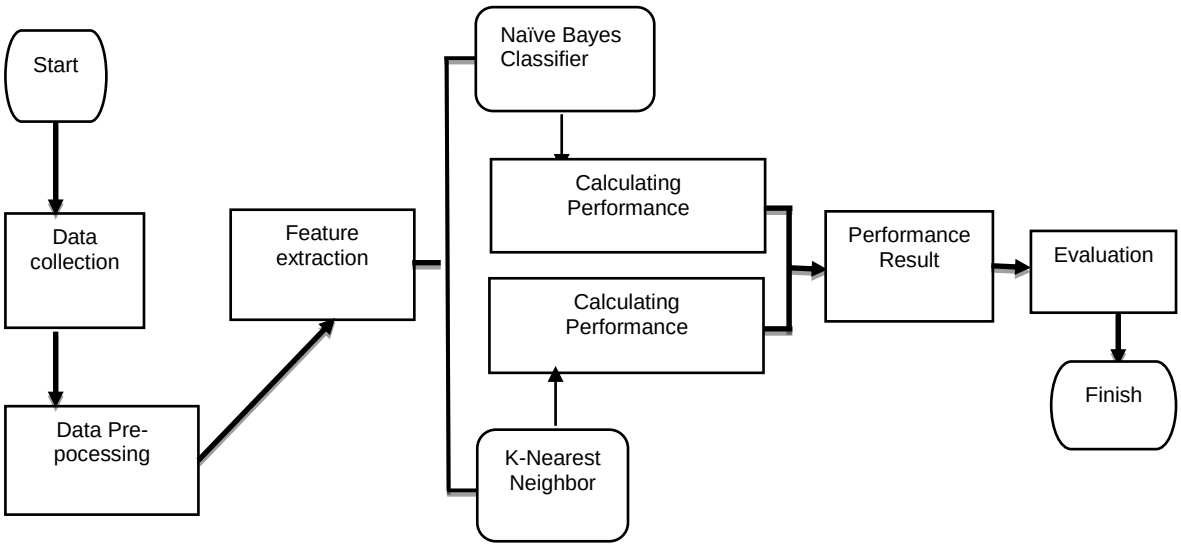


Figure 2. Algorithm implementation flow

Figure 2 is an algorithm implementation flow that explains the steps in completing the calculation of the two algorithms for herbal leaf classification and compares the performance of both algorithms.

1. Naïve Bayes Classifier

The Naïve Bayes Classifier algorithm is a classification method based on Bayes' theorem assuming independence between features (Shaban et al., 2021). The Naïve Bayes Classifier algorithm formula can be written in equation (1)

$$P(y|x_1,..,x_n) = \frac{P(x_1,...,x_n|y).P(y)}{P(x_1,...,x_n)} \quad (1)$$

Where, $P(y|x_1,\dots..,x_n)$ is the posterior probability of the class (target) y with the knowledge of the features $x_1,\dots..,x_n$. $P(x_1,\dots..,x_n|y)$ is the conditional probability of the features that the class gives y. $P(y)$ is the prior probability of the class y. $P(x_1,\dots..,x_n)$ is the total probability of the features $x_1,\dots..,x_n$. In implementation, We often ignore the denominator $P(x_1,\dots..,x_n)$ because it is constant for all classes and focuses on maximizing the numerator.

2. K-Nearest Neighbor (KNN)

The K-Nearest Neighbor (KNN) algorithm is a classification (or regression) method that works by comparing new instances to k Nearest instance in training dataset. The classification of new instances is determined based on the majority of classes from k The nearest instance. There is no specific "formula" for KNN because this method relies more on measuring the distance between instances (Siddalingappa & Kanagaraj, 2022). The distance that is often used is the euclidean distance, although it can also use other distances such as manhattan or minkowski.

Euclidean distance between two points p and q with dimensions n calculated by the formula as in equation (2)

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2} \quad (2)$$

Where, $d(p,q)$ is the distance between the training data to the test data, (q_i) that is train data to-i, (p_i) test data to-i.

Algorithm Configuration

In research on herbal leaf classification, the Naïve Bayes Classifier and K-Nearest Neighbor (KNN) algorithms were carefully configured to test their effectiveness. The data pre-processing process prepares leaf images for analysis, followed by feature extraction using the Gray Level Co-occurrence Matrix (GLCM) to capture textures (Vishnoi et al., 2022). Naïve Bayes is implemented utilizing Bayes' theorem, focusing on posterior probability, while KNN optimizes k selection to determine classes based on nearby instances (Banchhor & Srinivasu, 2021). Both are optimized through techniques such as cross-validation and evaluated with accuracy, precision, recall, and F1-score metrics, providing a comprehensive view of each other's performance in herbal leaf classification (El Akhal et al., 2023).

Model Evaluation

Model evaluation will use confusion metrics such as accuracy, precision, recall, and F1-score to assess the performance of both algorithms (Alem & Kumar, 2022). Statistical analysis may also be performed to determine significant differences between the performance of the two algorithms (Uddin et al., 2022). The formula for accuracy, precision, recall, and F1-score is written in equations (3) – (6)

$$Accuracy = \frac{(TP + TN)}{(TP + TN + FP + FN)} \quad (3)$$

Where TP is True Positive, that is, the number of positive samples that are correctly predicted. TN is True Negative, that is, the number of negative samples that are correctly predicted. FP is False Positive, i.e. the number of negative samples that are incorrectly predicted

as positive. And FN is False Negative, that is, the number of false positive samples is incorrectly predicted as negative (Hicks et al., 2022).

$$Precision = \frac{TP}{(TP + FP)} \quad (4)$$

Where TP is True Positive, that is, the number of positive samples correctly predicted by the model. FP is False Positive, i.e. the number of negative samples that are incorrectly predicted as positive by the model (Roy & Kumar, 2022).

$$Recall = \frac{TP}{(TP + Fn)} \quad (5)$$

Where TP is True Positive, that is, the number of positive samples correctly predicted by the model. FN is False Negative, i.e. the number of positive samples incorrectly predicted as negative by the model (Pellegrino et al., 2021).

$$F1\,Score = \frac{2x\,(precisiom\,x\,Recal)}{(2x(precisiom + Recal)} \quad (6)$$

3. RESULTS AND DISCUSSIONS

This study aimed to compare the effectiveness of two classification algorithms, Naïve Bayes and K-Nearest Neighbor (KNN), in classifying herbal leaves based on texture features extracted using the Gray Level Co-occurrence Matrix (GLCM). The dataset used consisted of images of papaya leaves, pandanus, cat's whiskers, and betel nut taken in different lighting conditions. This study emphasizes evaluating the performance of both algorithms using accuracy, precision, recall, and F1-score metrics.

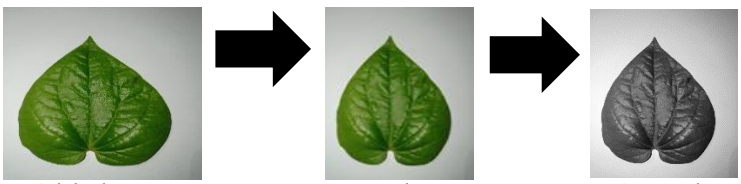
Data Collection

Table 1. Herbal leaf data

Types of Herbal Leaves	Picture	Information
Papaya leaf		50 Images
Pandan leaf		50 Images
Cat's whiskers Leaf		50 Images
Betel leaf		50 Images

The image data of herbal leaves in table 1 are papaya leaves, pandan leaves, cat's whisker leaves, and betel leaves. The data was taken from the results of portraits using the Redmi Note 9 mobile phone with a total of 200 images of herbal leaves where each type of leaf consists of 50 images. Table 1 shows the image of herbal leaf species in the dataset used.

Data Pre-processing



Original Image Resize Grayscale

Figure 3. Prepocessing stages

The leaf image prepocessing step applied in this study includes changing the image size from 2992 x 4000 pixels to 612 x 818 pixels, followed by conversion to grayscale, as shown in figure 3. The purpose of this preprocessing is to improve efficiency and accuracy in data analysis. Grayscale reduces the complexity of image data by eliminating color information that may not be relevant for a particular task, making it easier to extract features and their subsequent processing. This step can also reduce processing time and memory requirements, allowing algorithms to focus on important information on textures.

Feature Extraction

Feature extraction using the Grey Level Co-occurrence Matrix (GLCM) is a process that identifies image textures by analyzing pairs of pixels that have a specific intensity value. Features such as contrast, dissimilarity, homogeneity, and energy are calculated from the GLCM to provide detailed information about texture patterns in the image. This process helps in classification or pattern recognition by effectively distinguishing the texture characteristics of the image.

Table 2. Glcm calculation results

Contrast	Dissimilarity	Homogeneity	Energy	Label
40,18615921	2,634212222	0,541346424	0,09769697	Cat's whiskers
49,13527665	2,121270993	0,620747377	0,075958701	Cat's whiskers
62,20824813	3,126243002	0,512168047	0,081017727	Cat's whiskers
40,18615921	2,634212222	0,541346424	0,09769697	Cat's whiskers
62,20824813	3,126243002	0,512168047	0,081017727	Cat's whiskers
49,13527665	2,121270993	0,620747377	0,075958701	Cat's whiskers
60,38866502	3,215010864	0,4993331	0,075924782	Cat's whiskers
⋮	⋮	⋮	⋮	⋮
128,5686517	5,670440858	0,422666546	0,072298144	betel
74,534658	2,972262794	0,583332385	0,073048415	betel

The GLCM calculation results in table 2 show the values of texture features extracted from the leaf image, including contrast, dissimilarity, homogeneity, and energy, for each leaf label such as "pandan," "cat's whiskers," "betel," and "papaya." The contrast parameter measures the difference in intensity between pixels, dissimilarity calculates the absolute difference between pixels, homogeneity measures the proximity of the element distribution to the GLCM diagonal, and energy measures the smoothness of the image.

Naïve Bayes Results

Data calculation using Naïve Bayes yields:

Table 3. Naive bayes calculation results

No.	Types of leaves	Precision	Recall	F1-Score
1.	Cat's whiskers	0.62	0.83	0.71
2.	Pandan	0.90	0.60	0.72
3.	Papaya	0.90	1.00	0.95
4.	Betel	0.83	1.00	0.91
Overall Results		81.46%	85.83%	82.27%

Overall Acuracy = 82.50%

The calculation results using the Naïve Bayes algorithm for leaf type classification are shown in table 3. The Naïve Bayes algorithm is implemented assuming independence between features. The evaluation results showed that Naïve Bayes had an accuracy of 82.50%, with a precision of 81.46%, recall of 85.83%, and an F1-score of 82.27%. The main advantage of Naïve Bayes is its ability to handle high-dimensional datasets and provide fast results even on large data. However, the main drawback of this algorithm is the often unrealistic assumption of independence between features in leaf image data, which can result in decreased accuracy.

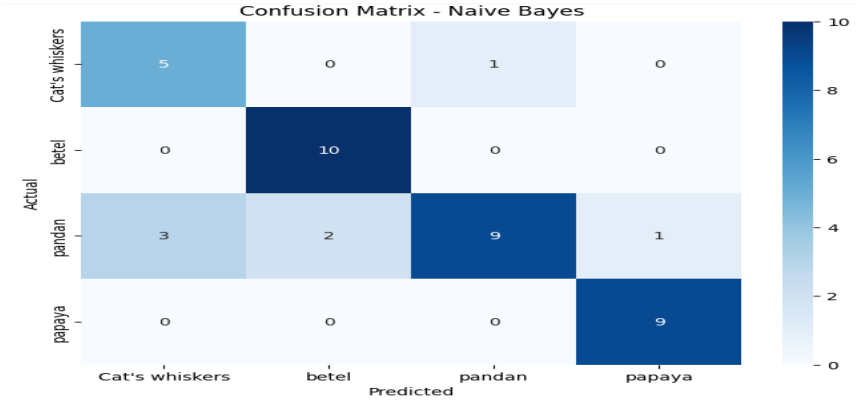


Figure 4. Naive bayes calculation chart

Figure 4 displays a confusion matrix for the results of the Naïve Bayes algorithm classification on leaf types. The row represents the actual class and the column represents the prediction class. The model managed to correctly classify the leaves of " Cat's whiskers " 5 times out of 6, "pandan" correctly 9 times out of 15, "papaya" correctly 9 times out of 9, and "betel" correctly 10 times out of 10. This matrix shows that the model has high accuracy for "papaya" and "betel," but some misclassification occurs in "cat's whiskers" and "pandanus."

K-Nearest Neighbor (KNN) Results

Data calculations using KNN produce:

Table 4. KNN calculation result

No.	Types of leaves	Precision	Recall	F1-Score
1.	Cat's whiskers	0.67	0.67	0.67
2.	Pandan	0.87	0.87	0.87
3.	Papaya	1.00	1.00	1.00
4.	Betel	1.00	1.00	1.00
Overall Results		88.33%	88.33%	88.33%

Overall Acuracy = 90.00%

The calculation results using the K-Nearest Neighbor (KNN) algorithm for leaf type classification are shown in Table 4. KNN is implemented by using euclidean distance to measure proximity between instances. The evaluation results show that KNN has a higher accuracy of 90.0%, with a precision of 88.33%, recall of 88.33%, and F1-score of 88.33%. The main advantages of KNN are its simplicity and ability to handle data with non-linear distribution without the need for data distribution assumptions. However, the disadvantage of KNN is that it requires higher computational time on large datasets, especially when determining the nearest neighbor.

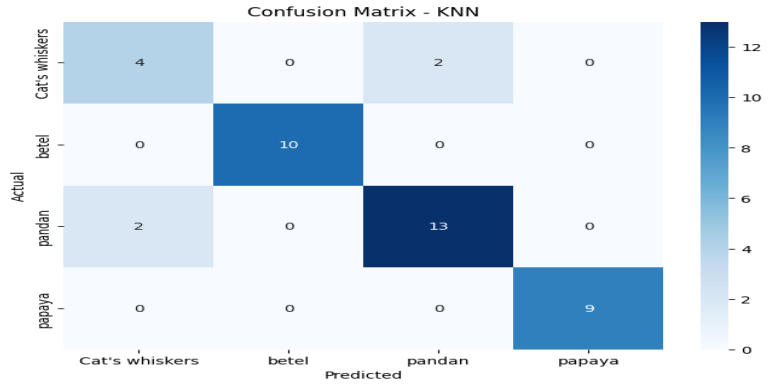


Figure 5. KNN calculation chart

Figure 5 shows a confusion matrix for leaf classification using K-Nearest Neighbor (KNN). The model managed to correctly classify the leaves of "cat's whiskers" 4 out of 6 times, but

incorrectly predicted 2 times as "pandan." The "pandan" leaf was predicted correctly 13 out of 15 times, with 2 mispredictions as "cat's whiskers." The leaves of "papaya" and "betel" were correctly classified 100% from 9 and 10 samples, respectively. This matrix indicates that KNN is very accurate for "papaya" and "betel" leaves, but there are some errors for "cat's" and "pandan" leaves.

Performance Result

Algorithms are evaluated based on accuracy, precision, recall, and F1-Score metrics

Table 5. Algorithm performance results

Algoritma	Acuracy	Precision	Recall	F1-Score
Naïve Bayes	82.50%	81.46%	85.83%	82.27%
KNN	90.00%	88.33%	88.33%	88.33%

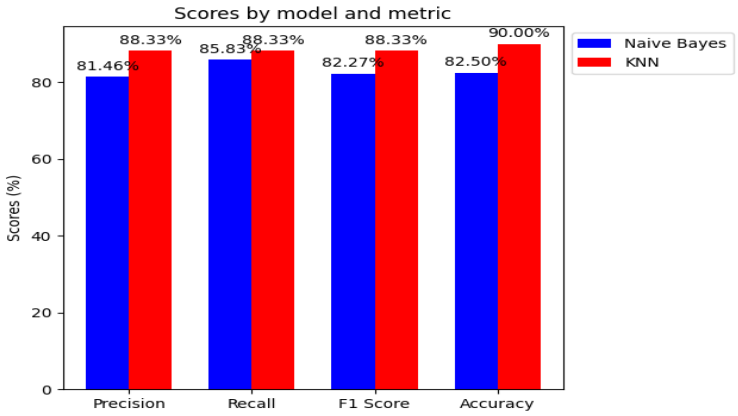


Figure 6. Comparison results graph

Table 5 and figure 6 show that K-Nearest Neighbor (KNN) is superior to Naïve Bayes in all evaluation metrics. KNN's precision is 88.33%, higher than Naïve Bayes' 81.46%. KNN recall was 88.33% compared to 85.83% for Naïve Bayes. KNN's F1-score is 88.33%, while Naïve Bayes is 82.27%. KNN's accuracy reaches 90.00%, better than Naïve Bayes' 82.50%. This graph confirms that KNN performs better in leaf type classification.

Evaluation

The calculation results using the Naïve Bayes algorithm show that this model has an overall accuracy of 82.50%, with a precision of 81.46%, recall of 85.83%, and an F1-score of 82.27%. Analysis of the confusion matrix showed that Naïve Bayes had some errors in classifying "cat's whiskers" and "pandan" leaves, but performed quite well for "papaya" and "betel" leaves. Naïve Bayes works by assuming independence between features, which is often unrealistic in leaf image data, which can reduce accuracy.

Conversely, the calculation results using K-Nearest Neighbor (KNN) show better performance. KNN achieved an overall accuracy of 90.00%, with a precision, recall, and F1-score of 88.33% each. The confusion matrix for KNN shows perfect accuracy in classifying "papaya" and "betel" leaves, but there are some errors in classifying "cat's whiskers" and "pandan" leaves. KNN works by measuring proximity between instances, which makes it possible to handle data with more complex distributions.

An overall performance evaluation shows that KNN is superior in all metrics to Naïve Bayes. The comparison graph shows that KNN's precision, recall, F1-score, and accuracy are all higher than Naïve Bayes. This suggests that KNN is more effective in classifying leaf types based on image texture. Although KNN requires longer computational time, especially on large datasets, its advantages in accuracy and consistency make it a better choice for this leaf type classification task.

The current research on classifying herbal leaves using Naïve Bayes and K-Nearest Neighbor (KNN) builds upon previous studies by incorporating modern feature extraction methods and performance metrics. Previous research, such as that by (Kamath, 2024), utilized algorithms like SVM, Naïve Bayes, KNN, and Decision Trees for categorizing Ayurvedic herbs with significant

results. Similarly, K. Pushpanathan et al. explored the use of SVM, Naïve Bayes, KNN, and CNN to improve the accuracy of medicinal plant identification. In contrast, the current study focuses specifically on herbal leaf classification using texture features extracted through the Gray Level Co-occurrence Matrix (GLCM) and evaluates the performance of Naïve Bayes and KNN in various lighting conditions, achieving higher accuracy for KNN at 90.00% compared to Naïve Bayes at 82.50%.

This study advances the field by addressing the challenges of high-dimensional data and feature similarity between species, which were significant hurdles in past research. For instance, (Sachar & Kumar, 2021) highlighted the importance of feature extraction and classification techniques using leaf images, while (Behera et al., 2021). demonstrated high accuracy in classifying papaya fruit ripeness using transfer learning models. The current research further refines these approaches by applying GLCM for texture feature extraction and systematically comparing the effectiveness of Naïve Bayes and KNN in handling the specific nuances of herbal leaf classification. This comparison provides deeper insights into the strengths and limitations of each algorithm in botanical contexts, contributing to the development of more robust and accurate automatic identification systems for herbal leaves.

4. CONCLUSION

This study demonstrated that the K-Nearest Neighbor (KNN) algorithm outperformed Naïve Bayes in classifying herbal leaves, achieving 90.00% accuracy compared to 82.50% for Naïve Bayes. KNN's strength is its ability to handle complex data without assuming feature independence, though it requires more computation time. The research suggests future work should optimize algorithm parameters, integrate deep learning, and test larger datasets. Practically, the high accuracy of KNN can aid in developing automated systems for identifying medicinal plants, benefiting botanists and researchers. The use of Gray Level Co-occurrence Matrix (GLCM) for texture features can also enhance plant identification in various settings.

REFERENCES

Abraham, E. J., & Kellogg, J. J. (2021). Chemometric-guided approaches for profiling and authenticating botanical materials. *Frontiers in Nutrition*, *8*, 780228. https://doi.org/https://doi.org/10.3389/fnut.2021.780228

Alem, A., & Kumar, S. (2022). Deep learning models performance evaluations for remote sensed image classification. *IEEE Access*, *10*, 111784–111793. https://doi.org/10.1109/ACCESS.2022.3215264

Banchhor, C., & Srinivasu, N. (2021). Analysis of Bayesian optimization algorithms for big data classification based on Map Reduce framework. *Journal of big data*, *8*(1), 81. https://doi.org/https://doi.org/10.1186/s40537-021-00464-4

Behera, S. K., Rath, A. K., & Sethy, P. K. (2021). Maturity status classification of papaya fruits based on machine learning and transfer learning approach. *Information Processing in Agriculture*, *8*(2), 244–250. https://doi.org/https://doi.org/10.1016/j.inpa.2020.05.003

Edison, H., Wang, X., & Conboy, K. (2021). Comparing methods for large-scale agile software development: A systematic literature review. *IEEE Transactions on Software Engineering*, *48*(8), 2709–2731. https://doi.org/10.1109/TSE.2021.3069039

El Akhal, H., Yahya, A. Ben, Moussa, N., & El Alaoui, A. E. B. (2023). A novel approach for image-based olive leaf diseases classification using a deep hybrid model. *Ecological Informatics*, *77*, 102276. https://doi.org/https://doi.org/10.1016/j.ecoinf.2023.102276

Falk, A., Becker, A., Dohmen, T., Huffman, D., & Sunde, U. (2023). The preference survey module: A validated instrument for measuring risk, time, and social preferences. *Management Science*, *69*(4), 1935–1950. https://doi.org/https://doi.org/10.1287/mnsc.2022.4455

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, *12*(1), 5979. https://doi.org/https://doi.org/10.1038/s41598-022-09954-8

Howes, M. R., Quave, C. L., Collemare, J., Tatsis, E. C., Twilley, D., Lulekal, E., Farlow, A., Li, L., Cazar, M., & Leaman, D. J. (2020). Molecules from nature: Reconciling biodiversity conservation and global healthcare imperatives for sustainable use of medicinal plants and fungi. *Plants, People, Planet*, *2*(5), 463–481. https://doi.org/https://doi.org/10.1002/ppp3.10138

Kamath, V. (2024). Assessing classification approaches for categorizing Ayurvedic herbs. *Multimedia Tools and Applications*, 1–25. https://doi.org/https://doi.org/10.1007/s11042-024-19061-7

Kolhar, S., & Jagtap, J. (2023). Plant trait estimation and classification studies in plant phenotyping using machine vision–A review. *Information Processing in Agriculture*, *10*(1), 114–135. https://doi.org/https://doi.org/10.1016/j.inpa.2021.02.006

Meng, J., You, X., Zhang, X., Shi, T., Zhang, L., Chen, X., Zhao, H., & Xu, M. (2023). Remote Sensing Application in Chinese Medicinal Plant Identification and Acreage Estimation—A Review. *Remote Sensing*, *15*(23), 5580. https://doi.org/https://doi.org/10.3390/rs15235580

Mulugeta, A. K., Sharma, D. P., & Mesfin, A. H. (2024). Deep learning for medicinal plant species classification and recognition: a systematic review. *Frontiers in Plant Science*, *14*, 1286088. https://doi.org/https://doi.org/10.3389/fpls.2023.1286088

Naeem, S., Ali, A., Chesneau, C., Tahir, M. H., Jamal, F., Sherwani, R. A. K., & Ul Hassan, M. (2021). The classification of medicinal plant leaves based on multispectral and texture feature using machine learning approach. *Agronomy*, *11*(2), 263. https://doi.org/https://doi.org/10.3390/agronomy11020263

Pellegrino, E., Jacques, C., Beaufils, N., Nanni, I., Carlioz, A., Metellus, P., & Ouafik, L. (2021). Machine learning random forest for predicting oncosomatic variant NGS analysis. *Scientific reports*, *11*(1), 21820. https://doi.org/https://doi.org/10.1038/s41598-021-01253-y

Pushpanathan, K., Hanafi, M., Mashohor, S., & Fazlil Ilahi, W. F. (2021). Machine learning in medicinal plants recognition: a review. *Artificial Intelligence Review*, *54*(1), 305–327. https://doi.org/https://doi.org/10.1007/s10462-020-09847-0

Roy, P. K., & Kumar, A. (2022). Early prediction of COVID-19 using ensemble of transfer learning. *Computers and Electrical Engineering*, *101*, 108018. https://doi.org/https://doi.org/10.1007/s13369-021-05879-y

Sachar, S., & Kumar, A. (2021). Survey of feature extraction and classification techniques to identify plant through leaves. *Expert Systems with Applications*, *167*, 114181. https://doi.org/https://doi.org/10.1016/j.eswa.2020.114181

Shaban, W. M., Rabie, A. H., Saleh, A. I., & Abo-Elsoud, M. A. (2021). Accurate detection of COVID-19 patients based on distance biased Naïve Bayes (DBNB) classification strategy. *Pattern Recognition*, *119*, 108110. https://doi.org/10.1016/j.patcog.2021.108110

Siddalingappa, R., & Kanagaraj, S. (2022). K-nearest-neighbor algorithm to predict the survival time and classification of various stages of oral cancer: a machine learning approach. *F1000Research*, *11*. https://doi.org/10.12688/f1000research.75469.2

Tangkiatkumjai, M., Boardman, H., & Walker, D.-M. (2020). Potential factors that influence usage of complementary and alternative medicine worldwide: a systematic review. *BMC complementary medicine and therapies*, *20*, 1–15. https://doi.org/https://doi.org/10.1186/s12906-020-03157-2

Twum, F., Missah, Y. M., Oppong, S. O., & Ussiph, N. (2022). Textural Analysis for Medicinal Plants Identification Using Log Gabor Filters. *IEEE Access*, *10*, 83204–83220. https://doi.org/10.1109/ACCESS.2022.3196788

Uddin, S., Haque, I., Lu, H., Moni, M. A., & Gide, E. (2022). Comparative performance analysis of K-nearest neighbour (KNN) algorithm and its different variants for disease prediction. *Scientific Reports*, *12*(1), 6256. https://doi.org/https://doi.org/10.1038/s41598-022-10358-x

Ullah, H., Ahmad, B., Sana, I., Sattar, A., Khan, A., Akbar, S., & Asghar, M. Z. (2021). Comparative study for machine learning classifier recommendation to predict political affiliation based on online reviews. *CAAI Transactions on Intelligence Technology*, *6*(3), 251–264. https://doi.org/10.1049/cit2.12046

Upton, R., David, B., Gafner, S., & Glasl, S. (2020). Botanical ingredient identification and quality assessment: strengths and limitations of analytical techniques. *Phytochemistry Reviews*, *19*(5), 1157–1177. https://doi.org/https://doi.org/10.1007/s11101-019-09625-z

Vishnoi, V. K., Kumar, K., & Kumar, B. (2022). A comprehensive study of feature extraction techniques for plant leaf disease detection. *Multimedia Tools and Applications*, *81*(1), 367–419. https://doi.org/https://doi.org/10.1007/s11042-021-11375-0

Wu, N., Crusiol, L. G. T., Liu, G., Wuyun, D., & Han, G. (2023). Comparing machine learning algorithms for pixel/object-based classifications of semi-arid grassland in northern China using multisource medium resolution imageries. *Remote Sensing*, *15*(3), 750. https://doi.org/https://doi.org/10.3390/rs15030750