Published *online* on the journal's web page : http://innovatics.unsil.ac.id

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

An Analysis of Performance Variation in Classification Methods on Handwritten Lontara Numerals

Faida Daeng Bustam[1], Purnawansyah[2], Huzain Azis[3].

[1]*Universitas Muslim Indonesia, Jln Urip sumoharjo, Makassar,90231, Indonesia*
[2]*Universitas Muslim Indonesia, Jln Urip sumoharjo, Makassar, 90231, Indonesia*
[3]*Universiti Kuala Lumpur, Kuala Lumpur, Malaysia, 50250, Malaysia*

[1]*13020200190@umi.ac.id,* [2]*purnawansyah@umi.ac.id ,* [3]*huzain.azis@s.unikl.edu.my*

ARTICLE INFORMATION

Article History:
Received: July 15, 2024
Last Revision: September 25, 2024
Published *Online*: September 30, 2024

KEYWORDS

Handwritten Lontara digits,
Classification methods,
Performance analysis,
Cross-validation
Feature extraction

CORRESPONDENCE

Phone: +62 811-4484-875
E-mail: *huzain.azis@s.unikl.edu.my*

ABSTRACT

This research explores the performance of several classification algorithms on handwritten Lontara digits, a script traditionally used by the Bugis and Makassar communities in South Sulawesi, Indonesia. The dataset comprises 10,890-digit samples, contributed by 99 individuals, and is categorized into 10 distinct classes corresponding to the digits 0-9. The classification methods evaluated in this study include K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and Nu-Support Vector Classifier (NuSVC). Cross-validation techniques are employed to evaluate the performance of these classifiers using standard metrics such as accuracy, precision, recall, and F1 score. The findings demonstrate varying levels of performance across the algorithms. Notably, GNB achieves the highest recall, indicating its ability to correctly identify positive samples, whereas KNN and NuSVC exhibit moderate effectiveness across other performance metrics. KNN shows potential with its simple yet robust approach to classifying complex datasets, while NuSVC demonstrates a balanced performance, particularly in precision. However, all classifiers face challenges in achieving optimal accuracy, particularly due to the complexity of the handwritten Lontara digits, which exhibit unique and intricate patterns. The study concludes by suggesting that further improvements can be achieved by refining feature extraction techniques and optimizing the classifiers used. Enhancing feature extraction could provide better representations of the Lontara digits, potentially leading to improved classification accuracy. Additionally, algorithm optimization and the exploration of more advanced classification methods could further enhance the overall performance. This research provides a foundation for the development of automated recognition systems for Lontara script, contributing to its preservation and modern use.

1. INTRODUCTION

Lontara is one of the traditional writing systems used by the Bugis and Makassar people in South Sulawesi, Indonesia. In a modern context, the use of technology to recognize handwritten characters in Lontara script has become an important focus [1]. Lontara script has distinctive characters that differentiate it from Latin letters or other scripts, making it not easy to recognize. By recognizing Lontara script, it can be easily identified. The challenges in recognizing Lontara numeral handwriting, such as variations in writing styles and shape distortions, can significantly impact classification performance.

Algorithms that fail to handle these variations effectively will experience reduced accuracy, making it difficult for the model to generalize well to new data. This can lead to misinterpretation of results and introduce bias in classification, ultimately reducing the system's effectiveness in real-world applications. Therefore, a more robust and comprehensive approach is urgently needed to effectively address the inherent diversity and variability

present within the dataset, ensuring improved accuracy, reliability, and inclusivity across various data points, categories, and dimensions. An example of handwritten Lontara numerals is shown in Figure 1.

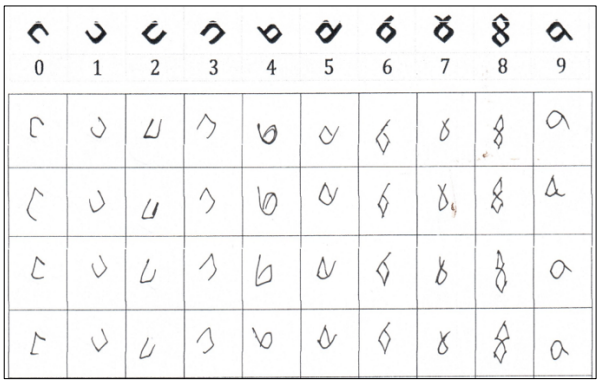

FIGURE 1. HANDWRITTEN LONTARA NUMERALS

Machine learning is a field in artificial intelligence that enables computers to learn from data and make decisions or predictions without being explicitly programmed [2], [3], [4], [5], [6], [7]. In machine learning, there are various types of algorithms used for specific tasks. One of them is the K-Nearest Neighbours (KNN) algorithm, which works by calculating the distance between a new data point and existing data points in the dataset, then selecting the class or target value based on the majority class of its K Nearest Neighbours. Another algorithm, Support Vector Machine (SVM), works by finding the best hyperplane that separates two classes in the given feature space, making it useful for classification and regression. Meanwhile, the Naïve Bayes algorithm is based on Bayes' theorem and assumes that features in the dataset are independent of each other, used for classification based on the probabilities of observed features [8], [9], [10], [11], [12], [13], [14], [15].

2. RELATED WORK

Previous research Implemented the Naïve Bayes Method for Handwritten Lontara Script Recognition [16]. This study comprised training and test data obtained directly from 5 respondents. The dataset collected is primary data because it is obtained directly from the respondents. The training data used consisted of 92 images of Lontara script obtained from 4 different individuals. The test data used consisted of 46 script images obtained from a different individual. The result of this study was an analysis of the accuracy of the Naive Bayes method in recognizing Lontara script images, achieving the best accuracy of 13.04%.

Research by Ahmad Angga et al. on the Implementation of Convolutional Neural Network (CNN) for Bima Script Handwriting Recognition This research aims to train a computer to recognize Bima script [17]. The Convolutional Neural Network (CNN) method is used in this study to recognize Bima script handwriting. The dataset used consists of 2640 images of Bima script handwriting with 22 classes. The results of the study show the reliable performance of the CNN model, with an accuracy of 97.34%, precision of 97.56%, recall of 97.34%, and an f1-score of 97.31% on the test data. Research by Ahlawat S and Choundary A on Hybrid CNN-SVM Classifier for Handwritten Digit Recognition [18]. In this research, a hybrid CNN-SVM model is proposed for handwritten digit recognition that involves automatic feature extraction using CNN and output prediction using SVM. This model combines the strengths of CNN and SVM classifiers in recognizing handwritten digits. The model also emphasizes the use of automatically generated features rather than hand-designed features. Experimental results show that their proposed approach achieves a classification accuracy of 99.28% for the MNIST dataset.

The research conducted by Anushka et al. "Cursive Handwriting Recognition Using CNN with VGG-16 [19] employs CNN with the VGG16 model to identify cursive English letters and words in specific scanned text documents. The first phase is image acquisition, which involves scanning image acquisition, image normalization, feature extraction from images, and segmentation application. Three different preprocessing techniques such as data augmentation, image segmentation, and image data generator were implemented in this work. CNN with the VGG16 model was used for recognition. Three types of experiments were conducted with various combinations of preprocessing techniques combined with the CNN model. The experimental results showed that the data augmentation preprocessing technique with CNN yielded the best performance, achieving a training accuracy of 98.36% and a testing accuracy of 95.1%.

Based on the provided research, the goal of the study is to evaluate and compare different classification methods for recognizing Lontara numeral handwriting. This involves examining the effectiveness of traditional methods like Naïve Bayes, which showed limited success with an accuracy of 13.04%, against more advanced techniques such as CNN, hybrid CNN-SVM, and CNN with VGG-16. Additionally, the study aims to identify the strengths and weaknesses of each method in handling complex numeral structures, improving accuracy, and reducing computational cost in practical applications related to Lontara script recognition.

The study aims to identify which method yields the highest accuracy and reliability, considering the significantly higher performance of CNN-based approaches in similar handwriting recognition tasks. Ultimately, the research seeks to recommend the most effective classification method for Lontara numeral handwriting, potentially incorporating data preprocessing techniques to enhance accuracy. This study provides insights into improving handwriting recognition systems, emphasizing modern techniques.

3. METHODOLOGY

The research process begins with data collection, where data is gathered either digitally or manually (e.g., handwriting on paper that is scanned). This is followed by core and stroke segmentation to process the handwriting data. Feature extraction is then performed to create the final dataset used in experiments. The dataset is tested using three different classification algorithms: K-Nearest Neighbor (KNN), Support Vector Machine (SVM), and Naive Bayes (NB). Each algorithm undergoes cross-validation to evaluate its performance, and a final analysis is conducted based on metrics such as accuracy and precision before drawing conclusions. The research stages are illustrated in Figure 2.

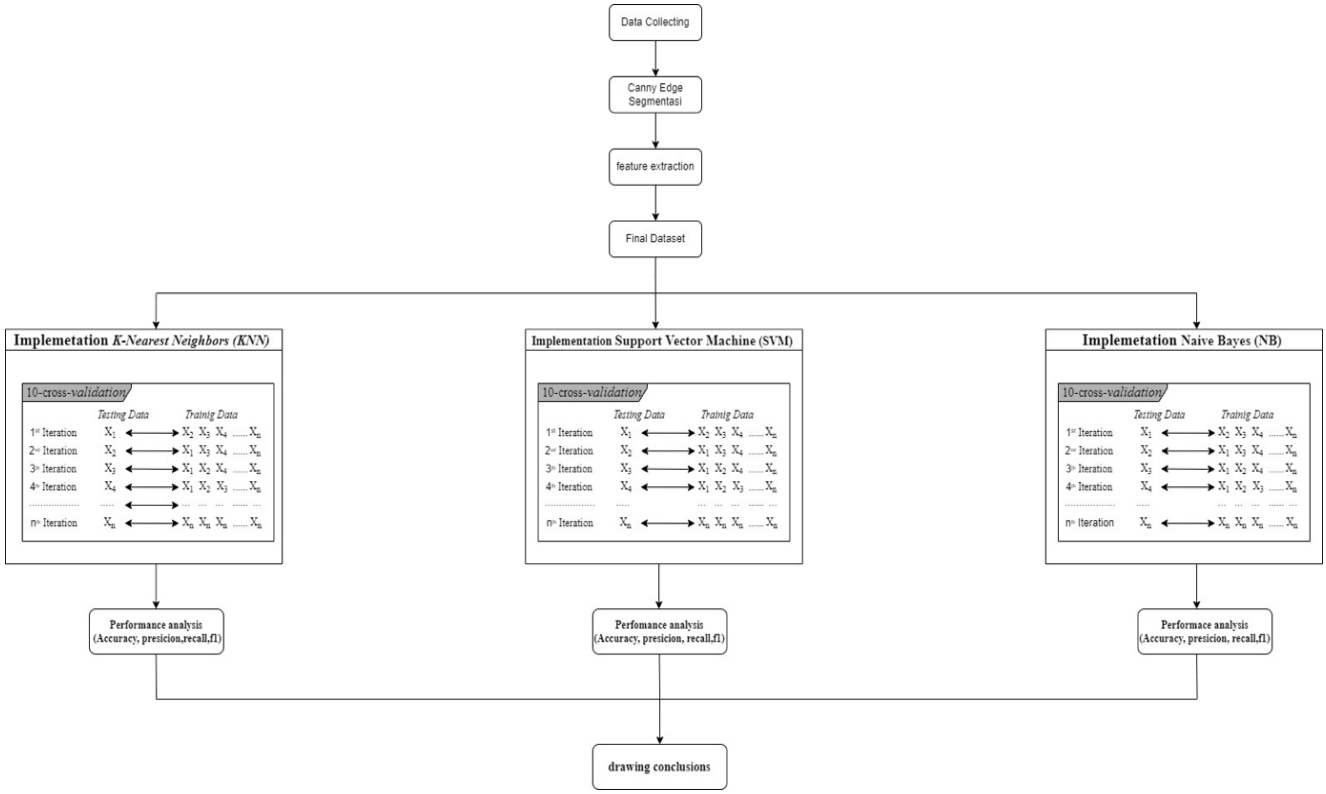

FIGURE 2. RESEARCH STAGES

3.1 Data Collection

Data collection can be carried out through various methods, including surveys, observations, interviews, and document analysis [20]. The data used in this study consists of 10,890 samples of handwritten Lontara numerals, collected from 99 writers with various levels of writing frequency and style variation. The writers consist of both females and males, each with different writing characteristics. The sample selection process is carefully carried out to cover a representative variation of Lontara numeral writing styles, ensuring that the research results reflect more realistic conditions [21].

FIGURE 3. LONTARA NUMERALS

3.2 Canny Edge Segmentation

Canny edge segmentation is an image processing technique used to detect edges in images [22], [23], [24], [25], [26]. The Canny algorithm works by identifying significant changes in pixel intensity between one area and its neighboring area, resulting in contours marking the edges of objects in the image. This process involves several steps, including applying a Gaussian filter to reduce noise, using gradient operators like Sobel to find image gradients, detecting edges by identifying pixels with the highest gradients, and finally applying hysteresis to distinguish true edges from false ones [27].

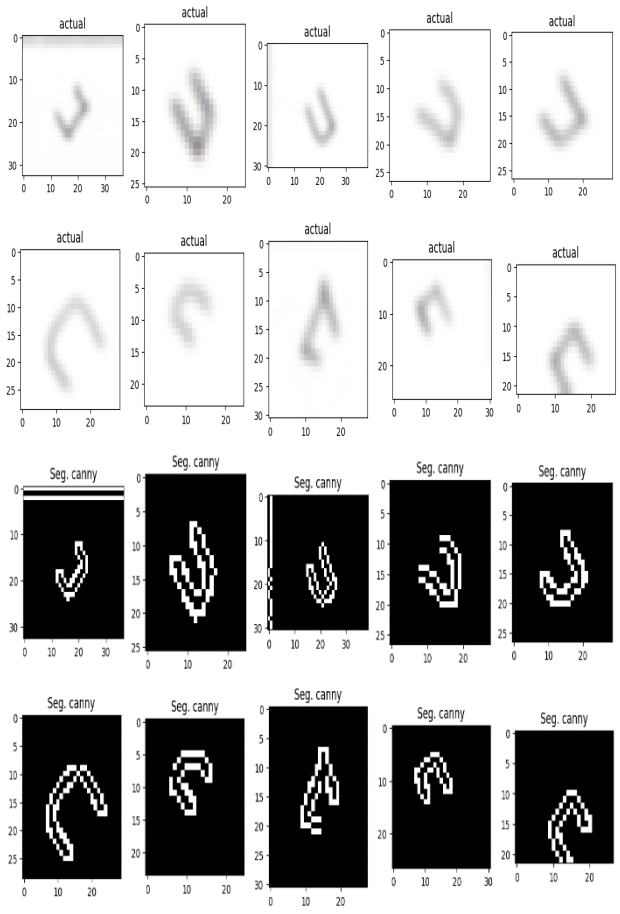

FIGURE 4. CANNY EDGE SEGMENTATION RESULT

3.3 Hu Moments Feature Extraction

Hu Moments is a feature extraction method used for object shape identification trough invariant moments [23], [28]. This technique has the main advantage of being able to capture numerical characteristics that do not change despite changes in rotation, scale, or translation of the object. In the context of Lontara script recognition, Hu Moments allows the system to recognize the shape of numbers despite variations in writing style or distortion caused by handwriting imperfections. The reliability of this method in detecting visual patterns makes it an effective tool for Lontara number recognition.

$$
\begin{aligned}
h_0 &= \eta_{20} + \eta_{02} \\
h_1 &= (\eta_{20} - \eta_{02})^2 + 4\eta_{11}^2 \\
h_2 &= (\eta_{30} - 3\eta_{12})^2 + (3\eta_{21} - \eta_{03})^2 \\
h_3 &= (\eta_{30} + 3\eta_{12})^2 + (\eta_{21} + \eta_{03})^2 \\
h_4 &= (\eta_{30} - 3\eta_{12})^2 + (\eta_{30} + \eta_{12})[(\eta_{30} - \eta_{12})^2 - \\
&\quad 3(\eta_{21} + \eta_{03})^2] + (3\eta_{21} - \eta_{03})[3(\eta_{30} + \\
&\quad \eta_{12})^2 - (\eta_{21} + \eta_{03})^2 \\
h_5 &= (\eta_{20} - \eta_{02})[\,(\eta_{30} + 3\eta_{12})^2 - (\eta_{21} + \\
&\quad \eta_{03})^2 + 4\eta_{11}(\eta_{30} + \eta_{12})(\eta_{21} + \eta_{03})] \\
h_6 &= (3\eta_{21} - \eta_{03}) + (\eta_{30} + \eta_{12})[(\eta_{30} - \\
&\quad \eta_{12})^2 - 3(\eta_{21} + \eta_{03})^2] + (\eta_{30} + \\
&\quad 3\eta_{12})(\eta_{21} + \eta_{03})[3(\eta_{30} + \eta_{12})^2 - \\
&\quad (\eta_{21} + \eta_{03})^2] \qquad (1)
\end{aligned}
$$

Hu Moments are essential in the efficient and robust analysis and recognition of images, as they capture shape-based features with high precision. The feature extraction process utilizing Hu Moments, as shown in Figure 5, not only highlights their ability to reduce dimensionality but also significantly enhances recognition accuracy. This is particularly beneficial in tasks like pattern recognition, image classification, and object detection, where accurate shape representation is critical. Furthermore, Hu Moments provide a compact yet comprehensive way of representing complex geometric shapes, making them indispensable for a wide range of computer vision applications, including image retrieval and automated visual systems.

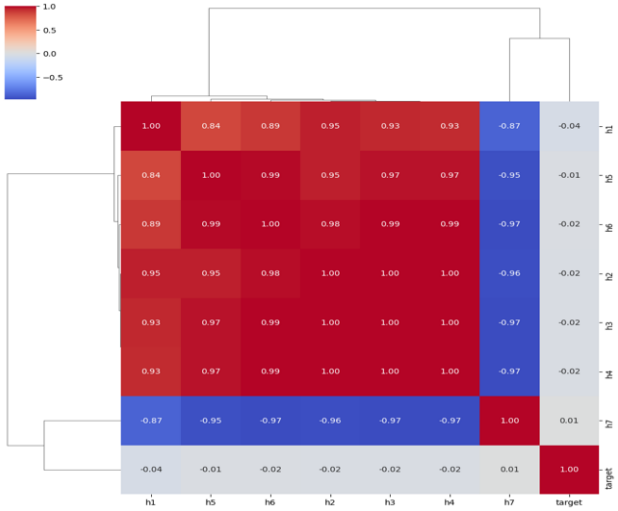

FIGURE 5. HU MOMENTS FEATURE EXTRACTION

3.4 Algorithm Implementation

In this study, we use libraries like scikit-learn to implement various classification models such as K-Nearest Neighbors (KNN), Gaussian Naïve Bayes (GNB), and Support Vector Machine (SVM). We also use Python as the main programming language for data processing and model implementation. Additionally, we will use statistical software such as NumPy and pandas for data analysis and visualization of results.

3.5 K-Nearest Neighbors

The K-Nearest Neighbors (KNN) algorithm is a widely used and fundamental classification technique in machine learning [29], [30]. This algorithm predicts the class of a data sample by examining the classes of its nearest neighbors in the feature space. It operates on the principle that similar data points are likely to belong to the same category. By considering the majority class among the nearest neighbors, KNN makes a classification decision. This simple yet effective approach is widely applied in various fields, including image recognition, recommendation systems, and medical diagnosis, due to its flexibility and ease of implementation.

$$d\,(x,y) = \sqrt{\sum_{i=1}^{n}(xi - yi)2} \qquad (2)$$

The K-Nearest Neighbors (KNN) algorithm aims to classify new objects by analyzing their attributes in relation to a set of labeled training samples. The core principle of KNN is to determine the nearest neighbors of a new data point by measuring the distance between it and the (K) closest points in the training dataset. The algorithm then assigns the class of most of these nearest neighbors to the new object. This approach relies on the assumption that data points with similar features tend to belong to the same category, making it effective for classification tasks. The distance between samples is calculated using the Euclidean distance formula [30], [31].

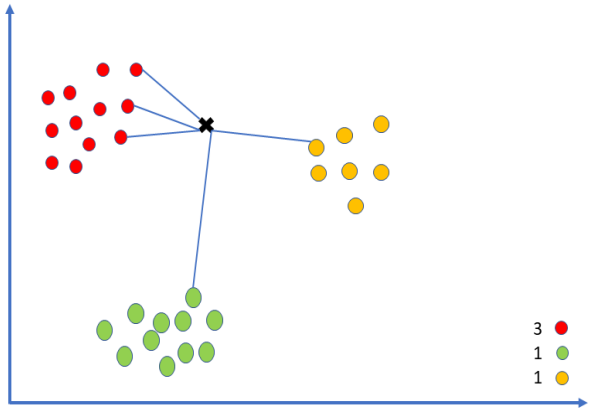

FIGURE 6. K-NEARST NEIGHBORS (KNN)

3.6 Support Vector Machine (SVM)

Support Vector Machine (SVM) is a machine learning algorithm used for classification and regression, aiming to find the optimal hyperplane that separates data into different classes[23], [32], [33]. This algorithm works by finding the maximum margin between data classes, so the nearest data points (support vectors) affect the hyperplane position. SVM can use kernel tricks to handle non-linearly separable data by mapping the data to a higher dimension. SVM is known for its ability to produce accurate and efficient results in various classification problems, although it can be slow on very large datasets. The SVM formula.

$$w.\ x + b = 0 \qquad (3)$$

Where:
w = weight vector
x = feature vector of data points
b = bias or intercept

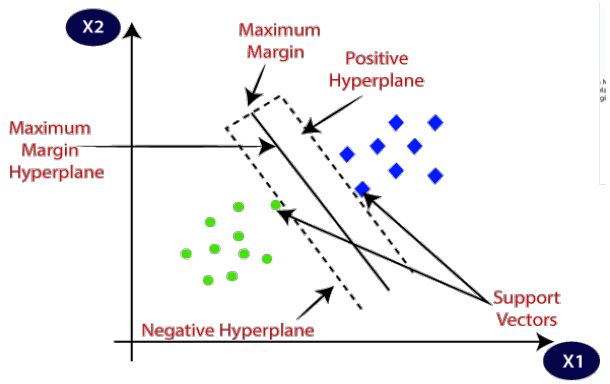

FIGURE 7. SUPPORT VECTOR MACHINE (SVM)

3.7 Naïve Bayes (NB)

Naïve Bayes (NB) is a machine learning classification technique that operates on the principles of Bayes' theorem. In this method, class predictions are made by calculating the probability of a specific class, given the observed features in the data. A key assumption of Naïve Bayes is that all features are independent of one another, simplifying the computation. Despite this assumption, the method often performs well in real-world applications, making it effective for tasks like text classification, spam detection, and sentiment analysis. The algorithm is particularly valued for its simplicity and efficiency in handling large datasets [34], [35], [36]. Bayes' theorem calculates probability of a hypothesis H given evidence E.

$$P(H \mid E) = \frac{P(E|H).\ P(H)}{P(E)} \qquad (4)$$

Where:
P (H | E) = posterior probability of hypothesis H given evidence E (probabilities posterior).
P (E | H) = likelihood of evidence E given hypothesis H (likelihood).
P (H) = prior probability of hypothesis H.
P (E) = marginal probability of evidence E

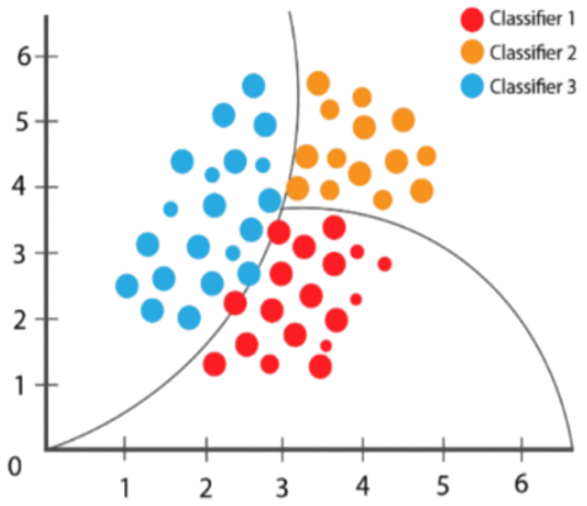

FIGURE 8. NAIVE BAYES (NB) CLASSIFIER

4. RESULT AND DISCUSSION

In this study, the balance of data in each numeric class (0-9) is an important consideration because it affects the classification results. If the amount of data in each class is unbalanced, the classification algorithm tends to focus more on the class that has more data, which results in bias in the results. Data imbalance can cause decreased performance, especially in class recognition with less data. Therefore, it is important to evaluate the data distribution to ensure that the model can recognize numbers with good accuracy in all classes. Gender or writing habits can affect the distribution and characteristics of data in handwriting recognition research. For example, differences in how men and women write, or how often a person is accustomed to writing, can create variations in the shape of the characters produced. Writing habits such as "first time," "rarely," or "often" can also produce different distortions in writing, which ultimately affect the recognition results. Therefore, these factors are important to consider in the analysis to minimize potential bias in the classification model.

Based on the results of classification using KNN, Gaussian Naive Bayes (GNB), and NU SVC methods with a 5-fold cross-validation (CV), a comprehensive performance comparison has been conducted for each model. The KNN model yielded an accuracy of 0.18, with precision, recall, and F1-score all balanced at 0.18. The GNB method outperformed the others, achieving an accuracy of 0.20 and a precision of 0.24, though its F1-score was slightly lower at 0.17. In contrast, the NU SVC model demonstrated the weakest performance, with an accuracy of 0.14 and an F1-score of 0.09, indicating its inefficiency with this dataset. These findings, shown in Table 1, highlight the varying effectiveness of these algorithms.

TABLE 1. PERFORMANCE COMPARISON RESULT CV 5

Classification Method CV 5 / Canny	Accuracy	Precision	Recall	F1 Score
KNN	0.18	0.18	0.18	0.18
GNB	0.20	0.24	0.20	0.17
NU SVC	0.14	0.15	0.14	0.09

Based on the classification results using Cross-Validation 10 (CV 10), the KNN method achieved an accuracy of 0.20 with precision, recall, and F1-score all balanced at 0.20. Gaussian Naive Bayes (GNB) has a slightly lower accuracy at 0.19, but its precision is higher at 0.26, although the F1-score drops to 0.16. NU SVC shows the lowest performance with an accuracy of 0.14 and an F1-score of only 0.07, indicating that this model is less than optimal in handling data at CV 10. Cross validation 5 is in table 2. Based on the classification results using 10-fold cross-validation (CV 10), the KNN method achieved an accuracy of 0.20, with precision, recall, and F1-score all balanced at 0.20, demonstrating consistent but moderate performance. Although Gaussian Naive Bayes (GNB) had a slightly lower accuracy of 0.19, it exhibited a higher precision of 0.26, though its F1-score dropped to 0.16, reflecting an imbalance between precision and overall effectiveness. In contrast, NU SVC performed the worst, with an accuracy of only 0.14 and a notably low F1-score of just 0.07, indicating its unsuitability for this dataset under CV 10. These findings underscore significant

differences in the models' capabilities, with each one performing differently based on various metrics. The results from cross-validation fold 5 are summarized in Table 2 for further comparison.

TABLE 2. PERFORMANCE COMPARISON RESULT CV 10

Classification Method CV 5 / Canny	Accuracy	Precision	Recall	F1 Score
KNN	0.20	0.20	0.20	0.20
GNB	0.19	0.26	0.19	0.16
NU SVC	0.14	0.06	0.14	0.07

5. CONCLUSTION

The results of this study demonstrate that the KNN algorithm offers the most consistent performance, achieving an accuracy and F1-score of 0.20, suggesting it handles balanced datasets effectively. However, KNN struggles when faced with more complex data patterns. In contrast, the Gaussian Naive Bayes (GNB) algorithm excels in precision, with a score of 0.26, but suffers from a lower F1-score, indicating a tendency toward misclassification. The NU SVC algorithm performed the poorest, with an F1-score of just 0.07, highlighting its weakness in managing data variation and complexity. These strengths and limitations provide valuable insights for future optimization and refinement efforts. The findings of this study indicate that while the KNN algorithm demonstrates stable performance with an accuracy and F1-score of 0.20, it faces challenges with more intricate data patterns. Gaussian Naive Bayes (GNB), with its superior precision (0.26), still struggles due to its lower F1-score, suggesting misclassification issues. NU SVC, with the lowest F1-score of 0.07, exhibits significant difficulty in handling complex and varied data. These observations are crucial for guiding future research, optimization, and potential algorithmic improvements, ensuring better handling of both balanced and more complex datasets in future implementations.

REFERENCES

[1] A. El-Sawy, M. Loey, and H. El-Bakry, "Arabic Handwritten Characters Recognition using Convolutional Neural Network," *2019 10th International Conference on Information and Communication Systems, ICICS 2019*, vol. 5, pp. 147–151, 2017, doi: 10.1109/IACS.2019.8809122.

[2] M. Dima Genemo, "Federated Learning for Bronchus Cancer Detection Using Tiny Machine Learning Edge Devices," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 64–69, Mar. 2024, doi: 10.56705/ijodas.v5i1.116.

[3] X. Wang, "Machine learning-enabled risk prediction of chronic obstructive pulmonary disease with unbalanced data," *Comput Methods Programs Biomed*, vol. 230, 2023, doi: 10.1016/j.cmpb.2023.107340.

[4] G. Kunapuli, *Ensemble Methods for Machine Learning*. Shelter Island, NY 11964: Manning Publications, 2023.

[5] A. Y. Kusdiyanto and Y. Pristyanto, "Machine Learning Models for Classifying Imbalanced Class Datasets Using Ensemble Learning," *2022 5th International Seminar on …*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/10052887/

[6] E. Elbasi *et al.*, "Crop Prediction Model Using Machine Learning Algorithms," *Applied Sciences (Switzerland)*, vol. 13, no. 16, 2023, doi: 10.3390/app13169288.

[7] R. C. Chen, C. Dewi, S. W. Huang, and R. E. Caraka, "Selecting critical features for data classification based on machine learning methods," *J Big Data*, vol. 7, no. 1, Dec. 2020, doi: 10.1186/s40537-020-00327-4.

[8] R. Singh, T. Ahmed, A. Kumar, A. K. Singh, and ..., "Imbalanced breast cancer classification using transfer learning," *… ACM transactions on …*, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9037082/

[9] Z. M. Zain, "Heterogeneous ensemble classifiers for Malay syllables classification," 2020. doi: 10.1063/5.0023094.

[10] L. Cheng, R. Guo, K. S. Candan, and H. Liu, "Representation learning for imbalanced cross-domain classification," *Proceedings of the 2020 SIAM …*, 2020, doi: 10.1137/1.9781611976236.54.

[11] Y. Liu, "Imbalanced dataset classification algorithm based on NDSVM," *J Phys Conf Ser*, 2021, doi: 10.1088/1742-6596/1871/1/012153.

[12] S. Keskin and O. Sevli, "Machine Learning Based Classification for Spam Detection," *Sakarya University Journal of Science*, vol. 28, no. 2, pp. 270–282, Apr. 2024, doi: 10.16984/saufenbilder.1264476.

[13] R. B. Hadiprakoso, W. R. Aditya, F. N. Pramitha, and P. Siber, "ANALISIS STATIS DETEKSI MALWARE ANDROID MENGGUNAKAN," vol. 5, no. 1, pp. 1–5, 2022.

[14] F. Alifiana, M. F. Asnawi, I. A. Ihsannudin, M. Alif, and M. Baihaqy, "ANALISIS SENTIMEN APLIKASI DUOLINGO MENGGUNAKAN ALGORITMA," vol. 13, no. 2, pp. 223–230, 2023.

[15] I. F. Hawari *et al.*, "PENGARUH TEKNIK OVERSAMPLING PADA ALGORITMA MACHINE LEARNING DALAM KLASIFIKASI BODY MASS INDEX (BMI)," *Jurnal Riset dan Aplikasi Matematika*, vol. 08, no. 01, pp. 51–68, 2024.

[16] I. I. Saputri, P. Purnawansyah, and ..., "Implementasi Metode Naïve Bayes Pada Pengenalan Tulisan Tangan Lontara," *Buletin Sistem Informasi …*, 2021, [Online]. Available: https://jurnal.fikom.umi.ac.id/index.php/BUSITI/article/view/845

[17] A. A. Handoko, M. A. Rosid, and U. Indahyanti, "Implementasi Convolutional Neural Network (CNN) Untuk Pengenalan Tulisan Tangan Aksara Bima," *SMATIKA JURNAL: STIKI …*, 2024, [Online]. Available: https://jurnal.stiki.ac.id/SMATIKA/article/view/1196

[18] S. Ahlawat and A. Choudhary, "Hybrid CNN-SVM classifier for handwritten digit recognition," *Procedia Comput Sci*, 2020, [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050920307754

[19] A. A. Rangari, S. Das, and D. Rajeswari, "Cursive handwriting recognition using CNN with VGG-16," *… International Conference on …*, 2023, [Online].

Available: https://ieeexplore.ieee.org/abstract/document/10083561/

[20] I. P. Adi Pratama, E. S. Jullev Atmadji, D. A. Purnamasar, and E. Faizal, "Evaluating the Performance of Voting Classifier in Multiclass Classification of Dry Bean Varieties," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 23–29, Mar. 2024, doi: 10.56705/ijodas.v5i1.124.

[21] H. Azis and F. D. Bustam, "Handwritten Lontara Numerals (0-9) Image Dataset," 2024, *Mendeley Data*.

[22] W. Ye, Y. Xia, and Q. Wang, "An Improved Canny Algorithm for Edge Detection," *Journal of Computational Information Systems*, vol. 75, pp. 1516–1523, 2011, doi: 10.1109/WCSE.2009.718.

[23] B. S. Waluyo Poetro, Eny Maria, H. Zein, E. Najwaini, and D. H. Zulfikar, "Advancements in Agricultural Automation: SVM Classifier with Hu Moments for Vegetable Identification," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 15–22, Mar. 2024, doi: 10.56705/ijodas.v5i1.123.

[24] A. P. Wibowo, M. Taruk, Thomas Edyson Tarigan, and M. Habibi, "Improving Mental Health Diagnostics through Advanced Algorithmic Models: A Case Study of Bipolar and Depressive Disorders," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 8–14, Mar. 2024, doi: 10.56705/ijodas.v5i1.122.

[25] D. Hanif Pristian, D. Iskandar Mulyana, E. Donaldo, and T. Informatika Sekolah Tinggi Ilmu Komputer Cipta Karya Informatika, "Klasifikasi Deteksi Hama pada Buah Mangga dengan Citra Digital Sistematic Literatur Review (SLR)."

[26] P. Citra, "KOMBINASI SOBEL , CANNY DAN OTSU UNTUK SEGMENTASI CITRA," vol. 13, no. 2, pp. 102–107, 2022.

[27] C. Digital, N. Filsa, and B. P. Adhi, "Kinerja Algoritma Canny untuk Mendeteksi Tepi dalam Mengidentifikasi Kinerja Algoritma Canny u ntuk Mendeteksi Tepi dalam Mengidentifikasi Tulisan pada Citra Digital Meme Avalaiable at :," no. June 2019, 2020, doi: 10.21009/pinter.3.1.8.

[28] B. S. Waluyo Poetro, Eny Maria, H. Zein, E. Najwaini, and D. H. Zulfikar, "Advancements in Agricultural Automation: SVM Classifier with Hu Moments for Vegetable Identification," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 15–22, Mar. 2024, doi: 10.56705/ijodas.v5i1.123.

[29] A. Sinra and Husni Angriani, "Automated Classification of COVID-19 Chest X-ray Images Using Ensemble Machine Learning Methods," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 45–53, Mar. 2024, doi: 10.56705/ijodas.v5i1.127.

[30] I. G. Iwan Sudipa, R. A. Azdy, I. Arfiani, N. M. Setiohardjo, and Sumiyatun, "Leveraging K-Nearest Neighbors for Enhanced Fruit Classification and Quality Assessment," *Indonesian Journal of Data and Science*, vol. 5, no. 1, pp. 30–36, Mar. 2024, doi: 10.56705/ijodas.v5i1.125.

[31] K. Moon and A. Jetawat, "Predicting Lung Cancer with K-Nearest Neighbors (KNN): A Computational Approach," *Indian J Sci Technol*, vol. 17, no. 21, pp. 2199–2206, 2024, doi: 10.17485/ijst/v17i21.1192.

[32] R. Obiedat, R. Qaddoura, A. Z. Ala'M, L. Al-Qaisi, and ..., "Sentiment analysis of customers' reviews using a hybrid evolutionary svm-based approach in an imbalanced data distribution," *IEEE …*, 2022, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9706209/

[33] B. Huang, Y. Zhu, Z. Wang, and Z. Fang, "Imbalanced data classification algorithm based on clustering and SVM," *Journal of Circuits, Systems and …*, 2021, doi: 10.1142/S0218126621500365.

[34] V. Metsis, I. Androutsopoulos, and G. Paliouras, "Spam filtering with Naive Bayes - Which Naive Bayes?," *3rd Conference on Email and Anti-Spam - Proceedings, CEAS 2006*, 2006.

[35] H. Zhang, "The optimality of Naive Bayes," *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2004*, vol. 2, pp. 562–567, 2004.

[36] Ericha Apriliyani and Y. Salim, "Analisis performa metode klasifikasi Naïve Bayes Classifier pada Unbalanced Dataset," *Indonesian Journal of Data and Science*, vol. 3, no. 2, pp. 47–54, 2022, doi: 10.56705/ijodas.v3i2.45.

AUTHORS

Faida Daeng Bustam

A dedicated Computer Science student at Universitas Muslim Indonesia (UMI), with a strong passion for technology, innovation, and problem-solving. Actively involved in both academic and extracurricular pursuits, including collaborative projects and leadership roles. Aiming to drive meaningful progress in the field of information technology by applying critical thinking and hands-on skills. Committed to leveraging knowledge, creativity, and adaptability to contribute to technological advancements and build a successful, impactful career in the rapidly evolving IT industry.

Puranawansyah

Academic and researcher at the Faculty of Computer Science, Universitas Muslim Indonesia (UMI). Extensive experience in information technology and computer systems. Research focus in informatics, particularly artificial intelligence (AI). Active in teaching and various scientific activities, as well as national and international seminars. Numerous publications in reputable scientific journals. Dedicated to advancing knowledge and technology to promote education and the information technology industry in Indonesia.

Huzain Azis

Student doctoral at Universiti Kuala Lumpur (UniKL). Faculty member and researcher at the Faculty of Computer Science, Universitas Muslim Indonesia (UMI). Expert in artificial intelligence (AI) and data science. Possesses substantial experience in AI technology development and

application. Engages actively in education, diverse research initiatives, and scientific conferences both domestically and internationally. Authored numerous articles in esteemed scientific journals. Dedicated to the progression of knowledge and technology to benefit education and the IT sector in Indonesia.