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

Innovation in Research of Informatics (INNOVATICS)

| ISSN (Online) 2656-8993 |

Comparative Performance Evaluation of Classification Methods for Arabic Numeral Handwritten Recognition

Intan Novita Saly[1], Purnawansyah[2], Huzain Azis[3]

[1,2,3]*Universitas Muslim Indonesia, Sulawesi Selatan, Makassar, 90231, Indonesia*
[3]*Universiti Kuala Lumpur, Kuala Lumpur, Malaysia, 50250, Malaysia*

[1]*13020200162@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 13, 2024
Published *Online*: September 30, 2024

KEYWORDS

Handwriting classification,
KNN algorithm,
Gaussian naive bayes,
NU SVC,
Optical Character Recognition (OCR)

CORRESPONDENCE

Phone: +62 811-4484-875

E-mail: huzain.azis@s.unikl.edu.my

ABSTRACT

This study aims to evaluate the performance of various classification methods in recognizing handwritten Arabic numerals, particularly the K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and NU Support Vector Classifier (NU SVC) algorithms. In this study, a dataset of handwritten Arabic numerals consisting of 9,350 samples with 10 different classes was used. The research process involved data collection, data labeling, dividing the dataset into training and testing data, implementing classification algorithms, and performance testing using cross-validation methods. The results showed that NU SVC had more stable performance with accuracy close to KNN, while GNB showed the lowest performance. The conclusion of this study emphasizes that the selection of algorithms and parameter optimization is crucial to improve the accuracy and efficiency of handwriting recognition systems. Support Vector Machine (SVM) based algorithms proved to be superior in handling complex classification tasks compared to GNB. This study provides significant contributions to the field of handwriting recognition, particularly in the context of Arabic numeral handwriting, and can serve as a reference for developers of optical character recognition (OCR) systems in the future. Future research is recommended to increase the variety of datasets and further explore parameter optimization and data preprocessing techniques to improve system accuracy.

1. INTRODUCTION

Handwriting has been a form of communication since ancient times and remains relevant in today's digital age [1], [2]. Arabic numerals are symbols representing the Hindu-Arabic numeral system used alongside the Arabic alphabet. Arabic numerals, consisting of digits 0 to 9, are widely used in numerous applications such as Optical Character Recognition (OCR), data entry automation, financial transactions, and document processing. Therefore, a deep understanding of Arabic numeral handwriting recognition is crucial for developing systems that can automatically recognize, classify, and interpret handwritten or scanned data efficiently and accurately across diverse platforms and languages. This capability enhances machine learning algorithms, improves automation accuracy, and facilitates seamless integration in sectors like banking, healthcare, education, and more. As a result, research in this area holds significant importance for technological advancement.

The main challenge addressed by this research is the accuracy in classifying Arabic numeral handwriting, considering the variation in writing styles and individual characteristics among different writers. Variability in handwriting is influenced by factors such as writing frequency, personal style preferences, age, and cultural differences, making it difficult to develop consistent and reliable classification algorithms. These variations pose significant challenges for Optical Character Recognition (OCR) technology, often affecting accuracy and reliability in recognizing and processing handwritten text. Additionally, variations in writing tools, speed, and environmental factors further complicate the classification process, requiring advanced machine learning models to overcome these hurdles in handwriting recognition systems. Hence, there is an urgent need to explore innovative approaches to improve the precision and efficiency of Arabic numeral classification, particularly focusing on handwritten forms. An example of Arabic numeral handwriting is shown in Figure 1.

FIGURE 1. ARABIC NUMERAL HANDWRITING

Machine learning is a branch of artificial intelligence that allows computers to learn from data and make predictions or decisions without being explicitly programmed [3], [4], [5]. Several machines learning algorithms, including Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Naive Bayes (NB), are widely used in classification tasks. SVM excels in high-dimensional data by identifying an optimal hyperplane to separate classes. KNN classifies data based on its nearest neighbors, while NB, a probabilistic classifier, assumes feature independence and is often applied in text classification and sentiment analysis. Each algorithm offers unique advantages, making them suitable for various applications, from image recognition to natural language processing, depending on the specific requirements of the dataset and task.

The purpose of this research is to evaluate the performance of various classification algorithms such as KNN, Naive Bayes, and other algorithms in recognizing handwritten Arabic numerals. The analysis is conducted by considering variations in writing styles, the writer's gender, and writing frequency to determine their impact on accuracy. Evaluation metrics such as accuracy, precision, recall, F1 score, and ROC AUC are used to compare the results of each method. The findings of this research are expected to provide insights into the best algorithms that can be used in classifying handwritten Arabic numerals as well as recommendations for further development.

2. RELATE RESEARCH

Previous research by N. Altwaijry et al. [6] explored the use of Convolutional Neural Networks (CNN) for classifying Arabic numeral handwriting. They used a dataset consisting of thousands of handwritten numeral images and reported significant accuracy compared to traditional methods like KNN and SVM. This research highlighted the advantages of CNN in recognizing complex patterns and writing style variations. Research by Albahli et al. [7] evaluated the performance of several classification algorithms such as K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machines (SVM) on a dataset of Arabic numeral handwriting. Their results indicated that SVM with Radial Basis Function (RBF) kernel provided higher accuracy compared to other methods. This research is crucial in understanding how different algorithms can be optimized for recognizing Arabic numeral handwriting.

Meanwhile, Shin et al [8] important Features Selection and Classification of Adult and Child from Handwriting Using Machine Learning Methods. In their research, they used text data and online handwriting patterns collected using a pen tablet system. They used SFFS (Sequential Forward Floating Selection) for feature selection and adopted two classification algorithms, RF (Random Forest) and SVM (Support Vector Machine), for classifying adults and children. They selected common features from the SFFS-RF and SFFS-SVM classifiers and then also applied RF and SVM classifiers to classify adults and children. For the handwriting text dataset, their proposed SFFS system with RF classifier achieved 93.5% accuracy for 18 features and 89.8% accuracy for 9 features in the handwriting pattern dataset. After identifying common features, SFFS-RF also produced classification accuracy of 91.5% and 87.7% for handwriting text and handwriting pattern datasets respectively.

Additionally, research by Syaifudin et al. talk about handwriting prediction using the support vector machine method in web-based applications [9].This research focuses on developing a handwriting recognition system using the Support Vector Machine (SVM) algorithm, tailored for web-based applications. The approach involves feature extraction as a critical step to describe and classify handwriting images accurately. The testing phase comprised 80 trials across 16 data points, where the model successfully predicted the handwriting patterns correctly in 55 cases while failing in 25. This led to an overall accuracy rate of 68.75%. Future improvements could involve enhancing feature extraction techniques, increasing the dataset size, or employing hybrid models to improve prediction accuracy for better real-world application.

This research aims to evaluate the performance of various classification algorithms in recognizing handwritten Arabic numerals, as demonstrated in previous studies that showed the effectiveness of CNN and SVM models for similar tasks. The focus is to compare the performance of KNN, GNB, and NU SVC algorithms to determine the most optimal algorithm in terms of accuracy and efficiency. Evaluation metrics such as accuracy, precision, recall, F1 score, and ROC AUC are used to compare the results of each method. The findings of this study are expected to serve as a reference for further implementation in optical character recognition (OCR) applications and other related technologies.

3. METHODOLOGY

The research design used in this study is experimental, focusing on the performance analysis of various classification methods on a dataset of Arabic numeral handwriting. To analyze the data, we will use the K-Nearest Neighbors (KNN), Support Vector Machines (SVM), and Naive Bayes Classifier (NB) algorithms in the field of Machine Learning. The research steps are illustrated in Figure 2.

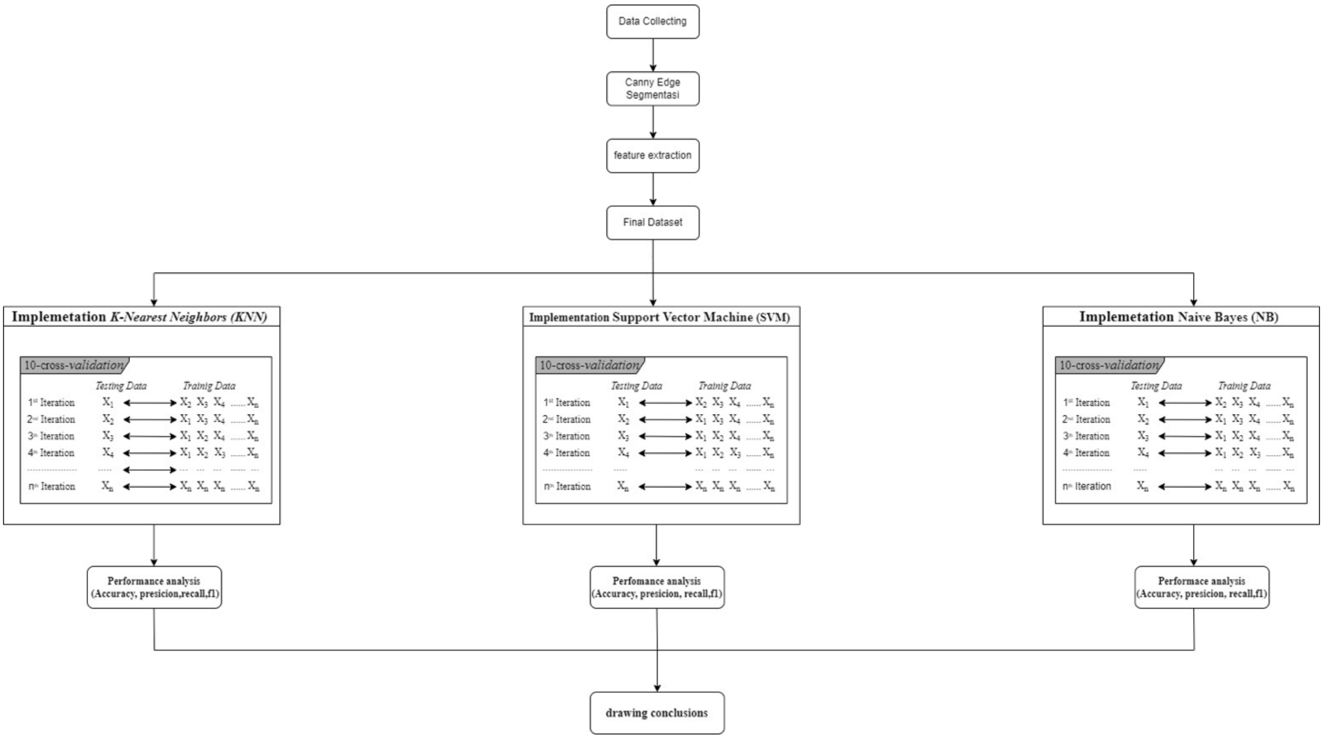

FIGURE 2. RESEARCH METHODOLOGY

3.1 Data Collection

Data collection is the process of gathering relevant information or data for a specific purpose [10],[11]. Data collection can be done through various methods, including surveys, observations, interviews, and document analysis. The data used in this research consists of 9,350 samples of Arabic numeral handwriting collected from 85 writers with varying writing frequencies and styles[12]. The writers include both men and women, each with different writing characteristics. The sample selection process was carefully conducted to include representative variations of Arabic numeral writing styles, so the research results can reflect more realistic conditions. The Arabic numerals example is shown in figure 3.

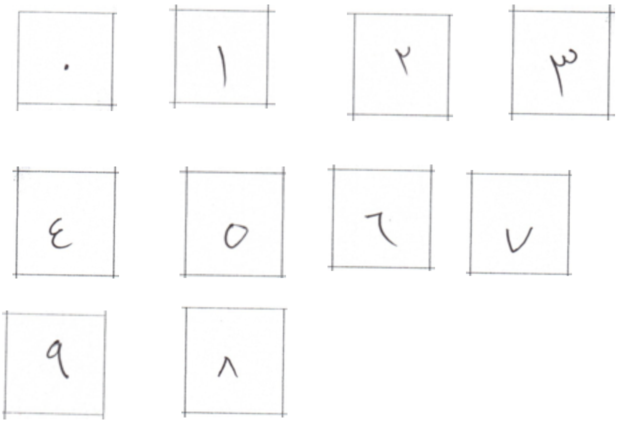

FIGURE 3. ARABIC NUMERAL

3.2 Canny Edge Segmentation

Canny edge segmentation is an image processing technique used to detect edges in images[13], [14], [15], [16]. The Canny algorithm works by identifying significant changes in pixel intensity between one area and its neighboring area, resulting in contours that mark object edges in the image. This process involves several steps, including applying a Gaussian filter to reduce noise, using gradient operators such as Sobel to find image gradients, detecting edges by identifying pixels with the greatest gradients, and finally applying hysteresis to determine true edges from false ones [17]. An example of Canny segmentation results is shown in Figure 4.

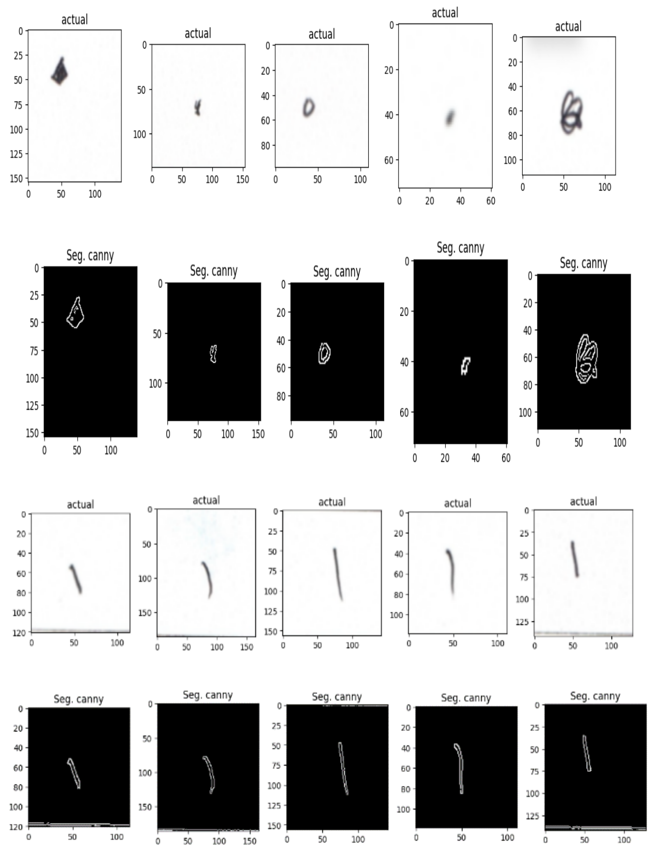

FIGURE 4. CANNY EDGE SEGMENTATION RESULT

3.3 Hu Moments Feature Extraction

Hu Moments feature extraction is a method in image processing used to capture shape information from objects in images [17], [18]. Hu Moments consist of seven invariant values derived from geometric moments and central moments of an image, meaning they remain unchanged even if the image undergoes transformations such as rotation, translation, or scaling. This makes Hu Moments very useful in pattern recognition and object classification, as these features can reliably recognize the same shape under different conditions. The extraction of Hu Moments begins by calculating the regular moments of the image, then using them to obtain central moments, and finally transforming these moments into seven Hu Moments. The following formula for Hu Moment can be translated into equation (1).

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

Thus, Hu Moments play a crucial role in analyzing and recognizing images more efficiently and robustly by capturing shape-based features. The Hu Moment feature extraction process is illustrated in Figure 5, demonstrating its effectiveness in reducing dimensionality and improving recognition accuracy, especially in pattern recognition and image classification tasks.

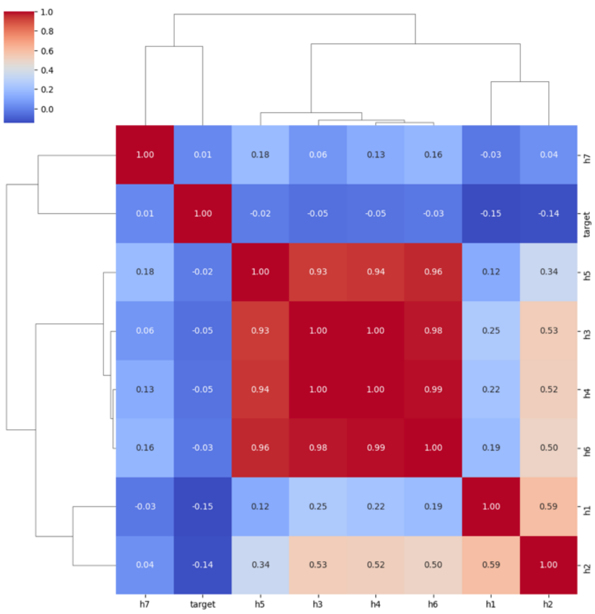

Figure 5. Hu Moments Feature Extraction

3.4 Algorithm Implementation

In this study, we will use several tools and technologies, including TensorFlow and Keras as frameworks to implement various classification models such as K-Nearest Neighbors (KNN), Gaussian Naive Bayes (GNB), and Support Vector Machines (SVM). We will also use Python as the main programming language for data processing and model implementation. Additionally, we will utilize statistical software such as NumPy and pandas for data analysis and result visualization. K-Nearest Neighbors (KNN).

3.5 K-Nearest Neighbors (KNN)

K-Nearest Neighbors (KNN) is a method that works by identifying the class or target value of a data point based on the majority class of its nearest neighbors in the feature space [19], [20]. KNN is non-parametric and lazy, meaning it does not require a stored model and only performs classification when new data appears. In other words, KNN predicts the label or value of new data based on the nearest data points in the given dataset [21], [22], [23]. The K-Nearest Neighbor (KNN) image is shown in Figure 6. The following formula for K-Nearest Neighbors (KNN) can be seen in equation (2):

$$d = \sqrt{(x_2 + x_1)^2 + (y_2 + y_1)^2} \quad (2)$$

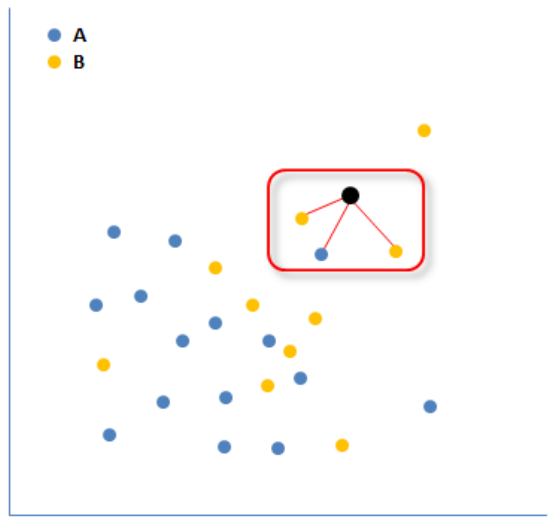

Figure 6. K-Nearst Neighboard (KNN)

3.6 Gaussian Naivest Bayes (GNB)

Gaussian Naive Bayes (GNB) is a classification algorithm based on Bayes' Theorem and the assumption that features in the data follow a normal (Gaussian) distribution[24], [25]. This algorithm calculates the probability that a data point belongs to a certain class based on feature values, assuming that each feature is independent of the others. GNB is highly effective and efficient in cases of normally distributed data and is often used for text classification and pattern recognition. With this probabilistic approach, GNB can handle data well despite imperfections and incompleteness, providing often accurate and quick classification results. The K-Nearest Gaussian Naive Bayes (GNB) image is shown in Figure 7. The following is the formula for Gaussian Naive Bayes (GNB) in equation (3) [24], [26], [27], [28].

$$P(X|Y) = \frac{P(Y|X) \cdot P(X)}{P(Y)} \quad (3)$$

Where:
X, Y = events
P (X|Y) = probability of X given Y
P (Y|X) = probability of Y given X
P(X), P(Y) = independent probabilities of X and Y

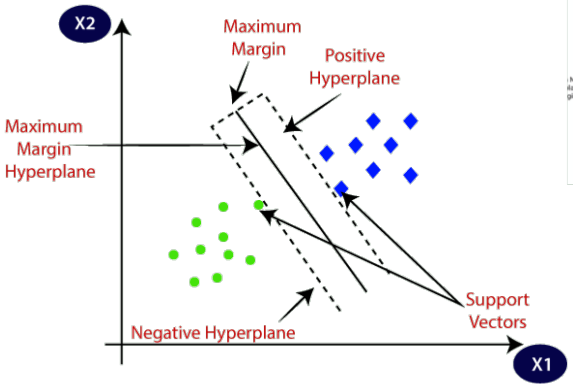

FIGURE 7. GAUSSIAN NAIVE BAYES

3.7 Support Vector Machines (SVM)

Support Vector Machine (SVM) is an algorithm in machine learning used for classification and regression tasks. The main goal of SVM is to find the best hyperplane that can separate two classes of data with maximum margin. In the context of classification, SVM works by finding a decision boundary that separates data groups into positive and negative classes. The hyperplane generated by SVM maximizes the distance between the two classes, thus improving prediction accuracy[10], [29], [30], [31], [32] [33], [34], [35]. The concept of SVM also involves "support vectors," which are data points closest to the separating hyperplane. SVM uses these support vectors to build an efficient and effective prediction model. Support Vector Machine (SVM) can be seen in Figure 8. The following formula for Support Vector Machine (SVM) can be seen in equation (4).

$$\omega \, . \, x + b = 0 \quad (4)$$

Where:
ω is the weight vector
x is the fearure vector
b is the bias

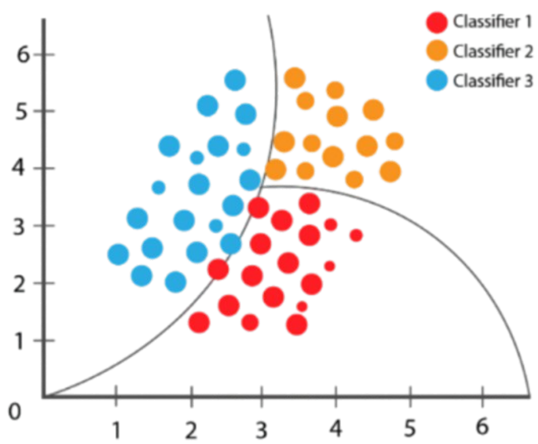

FIGURE 8. SUPPORT VECTOR MACHINE (SVM)

4. RESULT AND DISCUSSION

In this research, the data used is Arabic numeral handwriting consisting of 9,350 samples. This data covers 10 numeral classes (0-9) written by 33 women and 52 men with various levels of writing experience. This labeled data is then divided into training and testing datasets with a proportion of 80% for training and 20% for testing. All data is processed and prepared for use in training and testing various classification algorithms.

TABLE 1. PERFORMANCE COMPARISON RESULT CV 5

Classification Method CV 5 / Canny	Accuracy	Precision	Recall	F1 Score
KNN	0.33	0.33	0.33	0.33
GNB	0.16	0.24	0.16	0.10
NU SVC	0.34	0.32	0.34	0.32

TABLE 2. PERFORMANCE COMPARISON RESULT CV 10

Classification Method CV 5 / Canny	Accuracy	Precision	Recall	F1 Score
KNN	0.34	0.34	0.34	0.34
GNB	0.18	0.26	0.18	0.12
NU SVC	0.35	0.31	0.35	0.33

The results of the implementation of various classification algorithms are visualized in tables and graphs for easy analysis. The KNN performance table from 10-fold cross-validation shows variation in accuracy, precision, recall, and F1 score in each fold. The performance graph of NU SVC and GNB algorithms also shows performance metric comparisons across different cross-validation folds. These visualizations help understand and compare the effectiveness of each algorithm in recognizing Arabic numeral handwriting. The KNN results show an average accuracy of around 31.92%, indicating that this algorithm can recognize patterns in the data quite well. Conversely, the GNB performance with an average accuracy of 15.57% shows that this algorithm is less suitable for this classification task. NU SVC shows more stable performance with an average accuracy of 31.27%, indicating better potential compared to KNN and GNB for recognizing Arabic numeral handwriting.

A significant finding from this research is that NU SVC provides more stable results and approaches the performance of KNN, but there is still room for improvement. Although GNB shows low performance, it provides insights into the importance of selecting the right algorithm according to the data characteristics. This research highlights the need for further optimization in the use of algorithms to improve the accuracy and efficiency of Arabic numeral handwriting recognition systems. The research results show that the NU SVC algorithm has more stable performance, while GNB shows low performance, providing an important overview of the limitations of this algorithm in recognizing Arabic numeral handwriting. KNN with an average accuracy of around 31.92% shows that this algorithm is quite effective but still requires further optimization. Overall, the evaluation results show that selecting the right algorithm is very important to improve the accuracy of handwriting recognition systems.

The findings from this research are consistent with previous studies showing that SVM and its variations such as NU SVC often provide good performance in complex classification tasks. The relatively good results of KNN but requiring further optimization also align with literature stating that neighbor-based methods can work well if the right parameters are used. Conversely, the low performance of GNB confirms that the assumption of feature independence in this algorithm is often unmet in more complex datasets like handwriting. The research results show that NU SVC can be used as one of the classification methods in Arabic numeral handwriting recognition systems. However, additional steps such as parameter optimization and better data preprocessing techniques are needed to improve accuracy and other evaluation metrics. The use of ensemble methods or combinations with other techniques can also be considered to achieve better results.

One of the key limitations of this research lies in the restricted dataset variation, which may not fully capture the diverse range of Arabic numeral handwriting styles, potentially limiting the generalizability and robustness of the findings. Additionally, the absence of comprehensive parameter optimization for the tested algorithms, including KNN, GNB, and NU SVC, may have hindered the achievement of optimal performance results. These limitations suggest that the current results are not fully optimized and could benefit from further refinement. Future research should aim to enhance dataset diversity by incorporating a broader range of handwriting samples from writers of varied backgrounds. Additionally, exploring alternative algorithms and leveraging data augmentation techniques could significantly improve performance outcomes. Furthermore, detailed parameter optimization and advanced data preprocessing strategies should be considered to elevate the overall effectiveness of Arabic numeral handwriting recognition systems.

5. CONCLUSION

This research evaluates the performance of three classification algorithms KNN, GNB, and NU SVC in recognizing Arabic numeral handwriting. The findings reveal that NU SVC demonstrates more consistent performance, with accuracy comparable to KNN, while GNB lags significantly. This highlights the critical importance of selecting the appropriate algorithm and fine-tuning parameters to enhance both the accuracy and efficiency of handwriting recognition systems. The study identifies NU SVC as the most effective algorithm among those tested, affirming the hypothesis that SVM-based algorithms excel in handling complex classification tasks, while GNB is less suitable for data with dependent features. These results make a substantial contribution to the field of handwriting recognition by showing that NU SVC outperforms KNN and GNB in classifying Arabic numeral handwriting. Moreover, the research underscores the necessity of parameter optimization to further boost algorithmic performance. The insights from this study offer valuable guidance for developers of Optical Character Recognition (OCR) systems and other applications requiring accurate handwriting recognition. To enhance the validity of future results, researchers are encouraged to expand dataset diversity by including samples from a broader range of writers. Additionally, further efforts in parameter optimization, improved data preprocessing, and the exploration of ensemble methods or hybrid techniques could yield even higher performance in future studies.

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] N. Bonneton-Botté, L. Miramand, R. Bailly, and C. Pons, "Teaching and Rehabilitation of Handwriting for Children in the Digital Age: Issues and Challenges," Jul. 01, 2023, *Multidisciplinary Digital Publishing Institute (MDPI)*. doi: 10.3390/children10071096.

[3] F. T. Admojo and Y. I. Sulistya, "Analisis performa algoritma Stochastic Gradient Descent (SGD) dalam mengklasifikasi tahu berformalin," *Indonesian Journal of Data and Science (IJODAS)*, vol. 3, no. 1, 2022.

[4] A. A. SG Mas Karunia Maharani, K. Oka Saputra, and N. Made Ary Esta Dewi Wirastuti, "Comparison of Backpropagation Neural Network and Convolutional Neural Network on Handwriting Pattern Recognition". [Online]. Available: http://jcosine.if.unram.ac.id/

[5] M. Wibowo, M. Rizieq, and F. Djafar, "Jurnal Media Informatika Budidarma Perbandingan Metode Klasifikasi Untuk Deteksi Stress Pada Mahasiswa di Perguruan Tinggi," 2023, doi: 10.30865/mib.v7i1.5182.

[6] N. Altwaijry and I. Al-Turaiki, "Arabic handwriting recognition system using convolutional neural network," *Neural Comput Appl*, vol. 33, no. 7, pp. 2249–2261, Apr. 2021, doi: 10.1007/s00521-020-05070-8.

[7] N. Alrobah and S. Albahli, "Arabic handwritten recognition using deep learning: A survey," *Arab J Sci Eng*, 2022, doi: 10.1007/s13369-021-06363-3.

[8] J. Shin *et al.*, "Important Features Selection and Classification of Adult and Child from Handwriting Using Machine Learning Methods," *Applied Sciences (Switzerland)*, vol. 12, no. 10, May 2022, doi: 10.3390/app12105256.

[9] S. Syaifudin, D. Pratiwi, T. Rahardiansyah, "Handwriting prediction using the support vector machine method in web-based applications," *AIP Conference*, 2022, [Online]. Available: https://pubs.aip.org/aip/acp/article-abstract/2659/1/090004/2831896

[10] S. Kreatindo Manokwari, N. Yarkuran, I. Teknologi dan Bisnis Muhammadiyah Wakatobi, and F. Aziz, "Klasifikasi Kategori Obat Menggunakan Algoritma Support Vector Machine Drug Category Classification Using Support Vector Machine Algorithm," *JANUARI*, vol. 1, no. 1, p. 2023, 2023, [Online]. Available: https://www.kaggle.com/prathamtripathi/drug-

[11] N. Čerkez and V. Vareškić, "Machine Learning Approaches to Personality Classification on Imbalanced MBTI Datasets," *2021 44th International Convention*, 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9596742/

[12] H. Azis and I. N. Saly, "Handwritten Arabic Numerals (0-9) Image Dataset," 2024, *Mendeley Data*.

[13] H. Azis, N. Alfiyyah, and ; Purnawansyah, "Multiclass Classification on Nominal Value of Rupiah Banknotes Based on Image Processing," *ILKOM Scientific Journal*, vol. 16, no. 01, 2024, doi: 10.33096/ilkom.v1.

[14] H. Oumarou and N. Rismayanti, "Automated Classification of Empon Plants: A Comparative Study Using Hu Moments and K-NN Algorithm," *Indonesian*

Journal of Data and Science, vol. 4, 2023, doi: 10.56705/ijodas.

[15] 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.

[16] S. Niu, J. Yang, S. Wang, and G. Chen, "Improvement and parallel implementation of canny edge detection algorithm based on GPU," *Proceedings of International Conference on ASIC*, no. 6, pp. 641–644, 2011, doi: 10.1109/ASICON.2011.6157287.

[17] 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.

[18] Z. Huang and J. Leng, "Analysis of Hu's moment invariants on image scaling and rotation," in *Proc. of 2nd International Conference on Computer Engineering and Technology (ICCET*, 2010, pp. V7-476. doi: 10.1109/ICCET.2010.5485542.

[19] M. M. Baharuddin, T. Hasanuddin, and H. Azis, "Analisis Performa Metode K-Nearest Neighbor untuk Identifikasi Jenis Kaca," *Ilk. J. Ilm*, vol. 11, pp. 269–274, 2019.

[20] Y. Yang, H. G. Yeh, W. Zhang, C. J. Lee, and ..., "Feature extraction, selection, and K-nearest neighbors algorithm for shark behavior classification based on imbalanced dataset," *IEEE Sensors*, 2020, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9261357/

[21] S. Ferani *et al.*, "Klasifikasi Ulasan Aplikasi TikTok Menggunakan Algoritma K-Nearest Neighbor dan Chi Square."

[22] A. Huang, R. Xu, Y. Chen, and M. Guo, "Research on multi-label user classification of social media based on ML-KNN algorithm," *Technol Forecast Soc Change*, vol. 188, Mar. 2023, doi: 10.1016/j.techfore.2022.122271.

[23] K. Veena, K. Meena, Y. Teekaraman, R. Kuppusamy, and A. Radhakrishnan, "C SVM Classification and KNN Techniques for Cyber Crime Detection," *Wirel Commun Mob Comput*, vol. 2022, 2022, doi: 10.1155/2022/3640017.

[24] A. Nurul, Y. Salim, and H. Azis, "Analisis performa metode Gaussian Naïve Bayes untuk klasifikasi citra tulisan tangan karakter arab," *Indonesian Journal of Data and Science*, vol. 3, no. 3, pp. 115–121, 2022.

[25] Herman *et al.*, "Comparison of Artificial Neural Network and Gaussian Naïve Bayes in Recognition of Hand-Writing Number," in *2018 2nd East Indonesia Conference on Computer and Information Technology (EIConCIT)*, 2018, pp. 276–279. doi: 10.1109/EIConCIT.2018.8878651.

[26] A. Fitria and H. Azis, "Analisis Kinerja Sistem Klasifikasi Skripsi menggunakan Metode Naïve Bayes Classifier," *Prosiding Seminar Nasional Ilmu Komputer dan Teknologi Informasi*, vol. 3, no. 2, pp. 102–106, 2018.

[27] S. Sahar, "Analisis Perbandingan Metode K-Nearest Neighbor dan Naïve Bayes Clasiffier Pada Dataset Penyakit Jantung," *Indonesian Journal of Data and Science*, vol. 1, no. 3, pp. 79–86, 2020, doi: 10.33096/ijodas.v1i3.20.

[28] St. F. Fattah and Purnawansyah, "Analisis sentimen terhadap Body Shaming pada Twitter menggunakan Metode Naïve Bayes Classifier," *Indonesian Journal of Data and Science*, vol. 3, no. 2, pp. 61–71, 2022, doi: 10.56705/ijodas.v3i2.46.

[29] A. R. Damanik, S. Annisa, A. I. Rafeli, A. S. Liana, and D. S. Prasvita, "Klasifikasi Jenis Buah Cherry Menggunakan Support Vector Machine (SVM) Berdasarkan Tekstur dan Warna Citra," 2022. [Online]. Available: https://www.kaggle.com/moltean/fruits

[30] R. Oktafiani, A. Hermawan, and D. Avianto, "Pengaruh Komposisi Split data Terhadap Performa Klasifikasi Penyakit Kanker Payudara Menggunakan Algoritma Machine Learning," *Jurnal Sains dan Informatika*, pp. 19–28, Jun. 2023, doi: 10.34128/jsi.v9i1.622.

[31] J. Harwell and Y. Li, "Classifying Documents by Viewpoint Using Word2Vec and Support Vector Machines," in *Natural Language Processing and Information Systems*, P. Rosso, V. Basile, R. Martínez, E. Métais, and F. Meziane, Eds., Cham: Springer International Publishing, 2022, pp. 135–143.

[32] S. Rabbani, D. Safitri, N. Rahmadhani, A. A. F. Sani, and M. K. Anam, "Perbandingan Evaluasi Kernel SVM untuk Klasifikasi Sentimen dalam Analisis Kenaikan Harga BBM," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 3, no. 2, pp. 153–160, Oct. 2023, doi: 10.57152/malcom.v3i2.897.

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

[34] D. Datta, G. Singh, A. Routray, and ..., "Automatic Classification of Lithofacies with Highly Imbalanced Dataset Using Multistage SVM Classifier," *IECON 2021–47th*, 2021, [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9589254/

[35] R. Obiedat, R. Qaddoura, A. Z. Ala'M, L. Al-Qaisi, "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/

AUTHORS

Intan Novita Saly
Student at the Faculty of Computer Science, Universitas Muslim Indonesia (UMI). Passionate about technology and actively engaged in academic and extracurricular activities. Aims to contribute to advancements in information technology and pursue a successful career in the field. Hopes to make a significant impact in the growing field of AI and machine learning, while continuously improving technical skills and staying updated with emerging trends in the technology.

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.