Comparison of Similarity Distance-Based Metrics for HODA and BANGLA Dataset for Enhanced Precision

Mgd Maaz Taha Yassin., Amirul Ramzani Radzid., Mohd Sanusi Azmi., Nur Atikah Arbain

Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM), Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia

DOI: https://doi.org/10.47772/IJRISS.2025.91200013

Received: 09 December 2025; Accepted: 16 December 2025; Published: 30 December 2025

ABSTRACT

A similar metric is often used as a tool to measure the degree of similarity between two objects or pieces of data. It is essential in many areas of study including data analysis, machine learning and image processing, which provides a way to compare and evaluate the similarity of different entities. These metrics can be categorized into distance-based and similarity-based approaches, each with their strengths and applications. Therefore, this study is to do a comparison of various distance metrics on image classification performance using HODA and Bangla handwritten digit datasets. A comprehensive evaluation is conducted on eight different distance measures, namely Euclidean, Manhattan, Chebyshev, Canberra, Cosine, Minkowski, Jaccard, and Sorenson, within the Mean Average Precision (MAP) metric framework to evaluate their effectiveness in the context of handwritten digit recognition. Experimental results show that Chebyshev distance produces the highest classification accuracy of 71.6% on the HODA dataset, while Euclidean distance achieves the best performance on the Bangla dataset with 70.7% accuracy. In addition to quantitative analysis, a user study involving a structured questionnaire was conducted to qualitatively verify the MAP-based evaluation methodology. Results from user evaluations further reinforce the empirical findings. Therefore, the study underlines the importance of choosing an appropriate distance metric that is adapted to the specific properties of the dataset, highlighting its role in improving the performance of pattern recognition systems in computer vision applications.

Keywords: distance metric, image classification, HODA dataset, BANGLA dataset, MAP accuracy

INTRODUCTION

Distance metrics play a critical role in data analysis and pattern recognition, especially in image classification tasks, because they provide a quantitative way to measure how similar or dissimilar data points are. In the context of images, these metrics help compare feature vectors representations of visual characteristics extracted from images allowing algorithms to determine which images are most alike. By influencing how models cluster, classify, and retrieve images, distance metrics directly affect the accuracy and robustness of many computer vision systems. Choosing an appropriate metric can significantly enhance performance, as different metrics capture different aspects of variation within the data, such as pixel intensity differences, structural patterns, or high-dimensional feature relationships. The effectiveness of algorithms such as k-nearest neighbors (k-NN) and clustering depends on the selection of appropriate distance metrics to measure the similarities or differences between data points [1]. However, the performance of distance metrics often varies according to the data type, such as variations in resolution, texture, or image structure [2]. For example, Euclidean may be suitable for spherically distributed data, while Manhattan is more robust to outliers [3].

Handwritten datasets, such as Hoda and Bangla, are often used in this study because they present unique challenges due to variations in writing styles and noise [4]. These datasets exhibit significant diversity in character shapes, stroke patterns, and writing speeds, making them ideal benchmarks for evaluating the effectiveness of distance-based classification methods. Additionally, inconsistencies introduced during data acquisition such as blurred strokes, uneven illumination, and digitization artifacts further complicate the recognition process. These factors demand distance metrics that can robustly capture subtle similarities while remaining resilient to distortions and noise. Although traditional metrics such as Euclidean and Manhattan are commonly used, they have limitations, especially in dealing with complex variations in data features. This study and others have examined the comparison of various distance metrics, including Chebyshev, Euclidean, Manhattan, and others, in handwritten data classification. Mean Average Precision (MAP) is widely used as the main evaluation measure to measure classification effectiveness, providing a more accurate and holistic picture of performance than traditional measures. There are also innovative approaches that propose the use of MAP-based distance metrics to improve classification accuracy by overcoming the shortcomings of traditional metrics. The selection of distance metrics needs to be tailored to the nature and structure of the dataset used.

Incorporating distance metrics with text line segmentation will help in clustering and grouping the text contents. By leveraging these metrics, the system can more accurately determine spatial relationships between characters and words, ensuring that elements belonging to the same line are grouped together. This approach also helps minimize segmentation errors caused by irregular spacing, skewed writing, or overlapping strokes. Furthermore, using distance-based analysis allows the segmentation method to adapt dynamically to variations in handwritten or printed text layouts. The idea is to differentiate the distance between the lines of the text elements. This will allow the segmentation of the text lines is a document. This study was done by Amirul et. el. in 2015 for datasets Mushaf Al-Quran [5, 6]. Another study was done by using a method of text line segmentation using a hybrid projection based neighboring properties for Mushaf Al-Quran text [7]. Other than that, the segmentation by using multiphase level segmentation on Mushaf Al-Quran text also has been proposed [8]. By incorporating distance metrics with text segmentation or text analysis will help to achieve accurate and efficient segmentation.

Problem Statement

Data classification, particularly in the domain of digitized handwritten images such as the HODA and Bangla datasets, remains one of the central challenges in pattern recognition **[5]**. This difficulty arises from the high degree of variability inherent in human handwriting, including differences in stroke thickness, writing orientation, digit shape, and individual writing habits. Furthermore, the presence of noise, distortion, and uneven pixel distribution adds another layer of complexity. In such conditions, selecting an appropriate distance metric becomes essential, as it directly influences how similarity between samples is measured and, ultimately, how accurately a classifier can distinguish between different digit classes.

Traditional distance metrics such as Euclidean and Manhattan, although widely implemented in many recognition systems, often show inconsistent performance when applied to complex, high-variation datasets. These metrics assume a relatively uniform and linear distribution of data points, which is not always the case in real-world handwriting samples. As a result, their effectiveness decreases when confronted with nonlinear or irregular feature spaces. This limitation underscores the need for a more thorough evaluation of distance metrics beyond conventional approaches. To address this, the present study employs Mean Average Precision (MAP), a ranking-based evaluation method capable of assessing the entire retrieval list rather than only the top prediction. MAP provides a more comprehensive measure of classification effectiveness, especially in k-NN or similarity-based models.

Evidence from prior research further highlights the shortcomings of traditional metrics. Arbain et al. **[6]** demonstrated that the Euclidean distance can produce unstable classification results when dealing with nonlinear feature distributions, reinforcing concerns about its reliability across diverse datasets. This disparity between theoretical expectations and practical performance forms a central issue in the problem statement. It suggests that improving classification accuracy requires not only testing multiple distance metrics but also understanding their behavior in relation to specific dataset characteristics.

Thus, this study aims to identify and address the underlying causes of inaccuracy in distance-based classification of handwritten digits. It seeks to determine whether these inaccuracies stem from limitations in the selected similarity metrics, the nature of the datasets, or the ranking methodology used to assess performance. To strengthen this investigation, the report integrates insights from existing literature, offering a broader perspective on the challenges, methodological advancements, and evaluation strategies related to distance similarity measures and their assessment using MAP. By doing so, the study contributes to a deeper understanding of the relationship between distance metrics and classification outcomes, guiding future developments in pattern recognition and metric-based learning.

Objective

This study aims to conduct a comprehensive comparison of eight widely used distance metrics—Euclidean, Manhattan, Chebyshev, Canberra, Cosine, Minkowski, Jaccard, and Sørensen—in the context of handwritten digit classification for the HODA (Farsi) and Bangla datasets. By applying each metric within the same classification framework, the research evaluates their influence on recognition performance, robustness to noise, and sensitivity to variations in handwriting styles. This comparison provides insight into which metrics are better suited for particular data distributions and feature representations found in multilingual handwritten digit recognition tasks.

To enhance the evaluation process, the study introduces a Mean Average Precision (MAP)-based assessment method for measuring the ranking accuracy of classification results. Unlike traditional accuracy metrics that focus solely on whether the top prediction is correct, MAP evaluates the entire ranked list of retrieved neighbors. This offers a more detailed and informative measure of how effectively each distance metric distinguishes between similar and dissimilar samples. Incorporating MAP allows the study to capture the overall ranking behavior of each metric, making the evaluation more aligned with retrieval-based classification methods such as k-nearest neighbors.

Additionally, the study investigates the various factors that influence the effectiveness of these distance metrics based on the inherent characteristics of the datasets. Properties such as intra-class variability, feature dimensionality, image noise, stroke thickness, sparsity, and non-linear data distributions are examined to understand their impact on metric performance. By identifying which dataset features favor or hinder specific distance measures, the analysis provides deeper insight into why certain metrics outperform others in specific scenarios. This contributes to improved metric selection strategies and guides the development of more effective feature-distance combinations for future handwritten digit recognition systems.

LITERATURE REVIEW

Distance similarity algorithms constitute a fundamental approach for assessing the degree of resemblance between objects across a wide range of applications, including pattern recognition, information retrieval, and data classification [7]. In both image- and text-based pattern classification, an effective similarity measure is essential to accurately discriminate between objects that may exhibit subtle variations. Within the broader field of pattern recognition, the selection of an appropriate distance algorithm directly influences model performance, as the metric dictates how feature relationships are interpreted. Distance metric learning, in particular, aims to optimize similarity measurements based on dataset-specific characteristics, enabling more meaningful comparisons between objects and improving classification outcomes [8].

Handwritten data classification represents a challenging domain due to significant intra-class variability, differences in writing styles, and the presence of noise or distortion. These issues are well-documented in large handwriting datasets such as Bangla and Farsi, where writer diversity leads to inconsistent feature distributions [4], [5]. These factors often contribute to varying classification accuracy when different distance metrics are used. Consequently, determining an appropriate distance measure becomes a critical step in enhancing recognition accuracy for handwritten data. Prior studies have demonstrated that no single metric consistently outperforms others across all datasets; instead, performance varies depending on the structure and characteristics of the features being analyzed.

For instance, Dadang et al. [2] applied the Euclidean distance to classify Buni fruit based on shape and texture features, achieving an accuracy of 87%, which highlights Euclidean's effectiveness for structured, low-noise feature sets. Conversely, Suwanda et al. [3] reported that the Manhattan distance yielded superior clustering performance in the K-Means algorithm on the Iris dataset, demonstrating that Manhattan can be more robust than Euclidean in specific contexts. In biometric applications, Jaemin et al. [9] and Arnab et al. [15] found that distance-based methods applied to wavelet features can achieve competitive iris recognition accuracy, with certain metrics—such as Spearman or rank-based approaches—outperforming Euclidean and Cosine metrics depending on the feature distribution. These findings collectively underscore the dataset-dependent behavior of distance metrics.

Each distance metric carries inherent strengths and limitations. Euclidean distance is widely adopted due to its simplicity and computational efficiency; however, it is highly sensitive to outliers and may struggle with irregular feature spaces, as observed in plant and image classification tasks [12]. Chebyshev distance, which captures the maximum difference across dimensions, has been shown to be effective for datasets with substantial feature variation [13]. Meanwhile, Cosine similarity is advantageous for high-dimensional and sparse data, such as textual information, although its performance may degrade when applied to image datasets influenced by lighting variation or uneven pixel intensity [10], [14]. These differences emphasize the importance of systematic evaluation when selecting a metric for specific classification tasks.

Beyond the choice of distance metric, the evaluation methodology also plays a crucial role. Mean Average Precision (MAP) is widely recognized as an effective measure in information retrieval for assessing the quality of ranked outputs, including similarity-based retrieval systems [6]. When applied to classification, MAP provides a more nuanced evaluation by considering the ordering of predicted classes rather than only the top prediction. This is particularly valuable in situations involving class imbalance or where multiple candidate classes share similar features. Studies involving satellite image classification and metric-based models demonstrate that incorporating MAP yields a more robust and informative assessment of distance metric performance [16].

Table 1 provides a comprehensive overview of recent studies on distance similarity algorithms and their performance across various domains, including handwriting recognition, image classification, and biometric applications. The table highlights how different metrics such as Euclidean, Manhattan, Chebyshev, Cosine, and wavelet-based approaches perform differently depending on dataset characteristics and feature structures. For instance, Euclidean distance proved effective for structured, low-noise datasets like Buni fruit images [2] but was sensitive to outliers in more variable datasets such as medicinal plant images [12]. Manhattan distance was shown to outperform Euclidean in certain clustering tasks, such as K-Means on the Iris dataset [3], while Chebyshev distance worked well for datasets with substantial feature variation [13]. Cosine similarity excelled in high-dimensional sparse data but showed limitations in image datasets affected by lighting or pixel variation [10], [14]. Studies on iris recognition further confirm that the optimal metric is dataset-dependent, with rank-based metrics or wavelet-feature-based distances sometimes outperforming traditional measures [9], [15]. Additionally, evaluation methodologies like Mean Average Precision (MAP) enhance the robustness of similarity assessments by providing nuanced ranked evaluations rather than relying solely on top 1 accuracy [6], [16]. Overall, these findings underscore the critical importance of selecting appropriate distance metrics and evaluation methods tailored to specific datasets and applications.

Table 1: Summary of Literature on Distance Similarity Algorithms and Classification Performance

Study / Author	**Domain / Dataset**	**Distance Metric(s)**	**Key Findings**
Prapemrosesan et al. [7]	Student performance dataset	General distance matrices	Highlights importance of distance similarity algorithms in classification tasks
Alvarez-Melis & David [8]	Dataset similarity modeling	Optimal transport distance	Demonstrates improved similarity learning using geometric dataset distances
Mridha et al. [4]	Bangla handwriting dataset	-	Shows high intra-class variability affecting classification difficulty
Hossein & Ehsanollah [5]	Farsi handwritten digits	-	Demonstrates handwriting variation and challenges in pattern recognition
Dadang et al. [2]	Buni fruit images	Euclidean distance	Achieved 87% accuracy: effective for structured, low-noise features
Suwanda et al. [3]	Iris dataset (K-	Euclidean vs.	Manhattan outperforms Euclidean in

	Means)	Manhattan	certain clustering contexts
Jaemin et al. [9]	Iris recognition	Wavelet + distance-based metrics	Certain metrics outperform Euclidean/Cosine depending on feature structure

METHODOLOGY

This section presents a detailed description of the methodology employed in this research, highlighting the dataset, distance metrics, and evaluation framework used to analyze handwritten digit classification. The methodology is designed to ensure transparency, reproducibility, and a robust assessment of system performance across multiple similarity measures and evaluation phases.

Dataset

This study utilizes two widely recognized handwritten digit datasets to validate the proposed approach:

1. HODA Dataset: Comprises 60,000 images of Arabic handwritten digits, each with a resolution of 32×32 pixels. The dataset encompasses a wide range of handwriting styles, providing a challenging environment for evaluating classification performance.

2. Bangla Dataset: Contains 10,000 grayscale images of Bangla handwritten digits, representing diverse writing styles and intensities. The inclusion of both HODA and Bangla datasets ensures that the system is tested across multiple languages and scripts, enhancing the generalizability of the findings.

Both datasets undergo preprocessing to standardize image size, normalize pixel intensity, and reduce noise, which is essential for accurate feature extraction and subsequent similarity measurement

Distance Metrics

To evaluate similarity between images, this research compares eight distance metrics, each selected for its unique inductive bias and computational properties. The choice of metrics is motivated by their complementary strengths in handling different feature distributions and dataset characteristics:

1. **Euclidean Distance**: Measures the straight-line distance between two feature vectors in multidimensional space, widely used due to simplicity and efficiency.

$$\text{distance} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

It is used as a baseline distance metric because it is computationally efficient and well suited to continuous, isotropic feature spaces commonly found in image classification.

2. **Manhattan Distance:** Computes the sum of absolute differences across dimensions; often more robust than Euclidean when dealing with high-dimensional or sparse data.

$$\text{distance} = \sum_{i=1}^{n} |x_i - y_i|$$

It is more robust to outliers than Euclidean distance and is well suited to high-dimensional or sparse data where absolute differences are more informative.

3. **Chebyshev Distance**: Considers the maximum difference along any single dimension, making it effective for datasets with large feature variation.

$$\text{distance } = \max_{i=1}^{n}(|x_i - y_i|)$$

It captures the maximum deviation across dimensions, making it particularly effective for datasets with large stroke variations, such as HODA.

4. **Minkowski Distance** (p=3): A generalized form of Euclidean and Manhattan distances, providing flexibility to adjust the distance sensitivity through parameter p.

$$\text{distance } = \frac{1}{\sqrt[3]{\sum_{i=1}^{n} |x_i - y_i|^3}}$$

Generalizes L1 and L2 norms, parameter p controls gensitivity to large deviations, offering flexibility for intermediate behaviors.

5. **Jaccard Distance:** Measures dissimilarity between two sets by comparing shared versus unique elements, suitable for binary or sparse feature vectors.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

It is ideal for binary or set-based features; emphasize overlap versus uniqueness.

6. **Cosine Distance:** Evaluates the angular difference between vectors, effective for high-dimensional, sparse, or normalized data.

$$similarity(A,B) \ = cos(\theta) \ = \ \frac{A \cdot B}{||A||\ ||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2 \sum_{i=1}^{n} B_i^2}}.$$

Effective for high-dimensional, normalized data; focuses on orientation rather than magnitude, useful for sparse representations.

7. **Canberra Distance:** Emphasizes relative differences between feature values, giving more weight to dimensions with smaller magnitudes.

$$d(\mathbf{p},\mathbf{q}) = \sum_{i=1}^{n} \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

where

$$\mathbf{p} = (p_1, p_2, \ldots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \ldots, q_n)$$

It gives more weight to small-magnitude differences, making it sensitive to subtle variations in low-intensity features.

8. **Sorenson Distance:** Also known as the Dice coefficient, it considers the cardinalities of sets to measure similarity; particularly useful for set-based feature representations.

$$DSC = \frac{2|X \cap Y|}{|X| + |Y|}$$

where |X| and |Y| are the cardinalities of the two sets.

It is suitable for set-based or binary features; emphasizes proportional similarity.

Evaluation Framework Phase

In Pre-processing phase, the image normalization and feature extraction using Symlet wavelet. In Classification phase, the k-NN algorithm (k=5k=5) is used to classify images. For MAP Evaluation, accuracy is measured based on the ranking of classification results. Lastly, for User Testing phase, a questionnaire was conducted on 30 students to evaluate the usability of the system.

The selected distance metrics span norm-based, angular, and set-theoretic formulations to ensure robustness across diverse handwriting styles and structural variations. The Minkowski distance parameter was set to p=3 to interpolate between Manhattan (p=1) and Euclidean (p=2) distances, providing a balanced sensitivity to feature magnitude and noise. For feature extraction, Symlet wavelets were employed due to their near-symmetric properties and their effectiveness in capturing localized stroke transitions, which complement distance-based similarity comparisons in handwritten digit recognition.

To strengthen empirical evaluation and support performance comparisons, multiple complementary analyses were incorporated. Confusion matrices were reported for each distance metric and dataset to reveal class-specific error patterns. Precision–recall curves, including macro- and micro-averaged AUC-PR scores, were used to assess ranking performance beyond MAP. Additionally, statistical significance was evaluated using paired t-tests or Wilcoxon signed-rank tests on fold-wise MAP scores, with effect sizes reported and multiple-comparison corrections applied using the Holm–Bonferroni procedure.

RESULTS AND DISCUSSION

In this section, detailed explanations are provided regarding both the Distance Metric Performance and the Questionnaire Analysis. The discussion encompasses the methods used to evaluate each distance metric, the comparative findings that highlight their strengths and limitations, and an in-depth analysis of the questionnaire results that reflect user interactions and satisfaction. By integrating these two analytical components, this section aims to present a holistic understanding of the system’s technical performance and user-centered evaluation.

Evaluation Framework Phase

On the HODA dataset, Chebyshev achieved the highest accuracy (71.6%), followed by Euclidean (70.7%) and Manhattan (67.1%). This is due to Chebyshev’s ability to handle the high feature variation in Arabic images (Table 2). On the other hand, Euclidean performed better on the Bangla dataset (70.7%) due to the more uniform digit structure.

Table 2: Comparison of distance metric accuracy using MAP on dataset HODA and Bangla.

Distance Metric	Average MAP Accuracy (%)
Euclidean	70.7
Manhattan	67.1
Chebyshev	71.6
Canberra	60.0
Cosine	67.3
Minkowski	68.1
Jaccard	51.4
Sorenson	65.4

Table 2 reports MAP accuracy across metrics. Chebyshev attains the highest MAP for Euclidean and Minkowski follow closely. Figure 1 visualizes these differences.

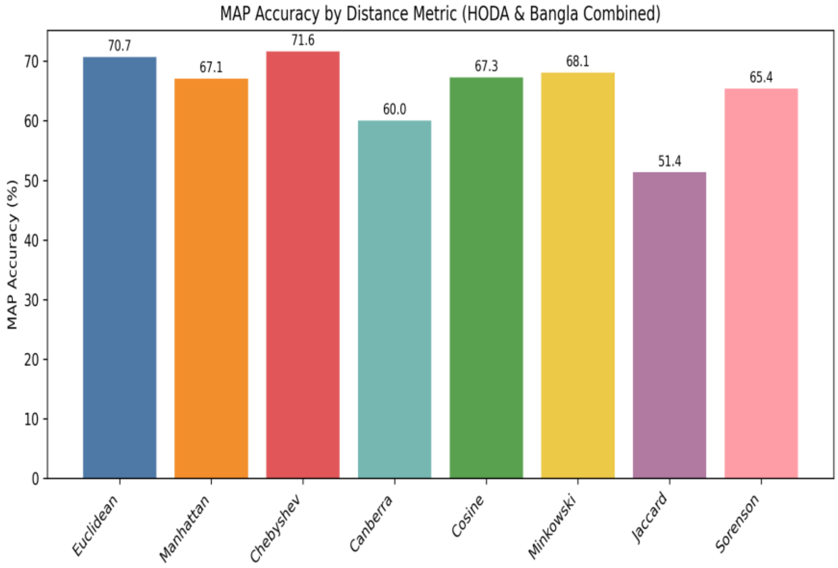

Figure 1. MAP accuracy by metric on HODA & Bangla.

Questionnaire Analysis

To evaluate user perceptions of the proposed system, this study employs a structured questionnaire comprising several key items. These items are designed to assess the perceived ease of integration of the MAP-based evaluation, the effectiveness of Chebyshev distance relative to Euclidean distance, the interpretability of system outputs, and the overall suitability of the system for handwritten digit recognition tasks.

Sample Questionnaire Items:

Q1. The MAP-based evaluation approach is easy to integrate into our existing workflow.

(1 = Strongly disagree … 5 = Strongly agree)

Q2. Using Chebyshev distance improves classification accuracy compared to Euclidean distance.

(1 = Strongly disagree … 5 = Strongly agree)

Q3. The results produced by the system are easy to interpret.

(1 = Strongly disagree … 5 = Strongly agree)

Q4. I would recommend this system for handwritten digit recognition tasks.

(1 = Strongly disagree … 5 = Strongly agree)

Participants: target N=30 (students). Record demographics: age range, gender, program/major, prior exposure to pattern recognition. Instruments: Likert-scale items (1–5) covering ease of integration, perceived accuracy, learnability, and trust; include open-ended questions for qualitative feedback. Reliability: compute Cronbach’s alpha; perform item–total correlation; remove items with low discrimination. Analysis: descriptive statistics (mean, SD, median, IQR), inferential tests (Mann–Whitney/Kruskal–Wallis for subgroup comparisons), and thematic coding for qualitative responses.

A total of 80% of respondents stated that MAP-based metrics are easy to integrate with existing systems (Figure 2). In addition, 75% agreed that Chebyshev improves classification accuracy compared to Euclidean.

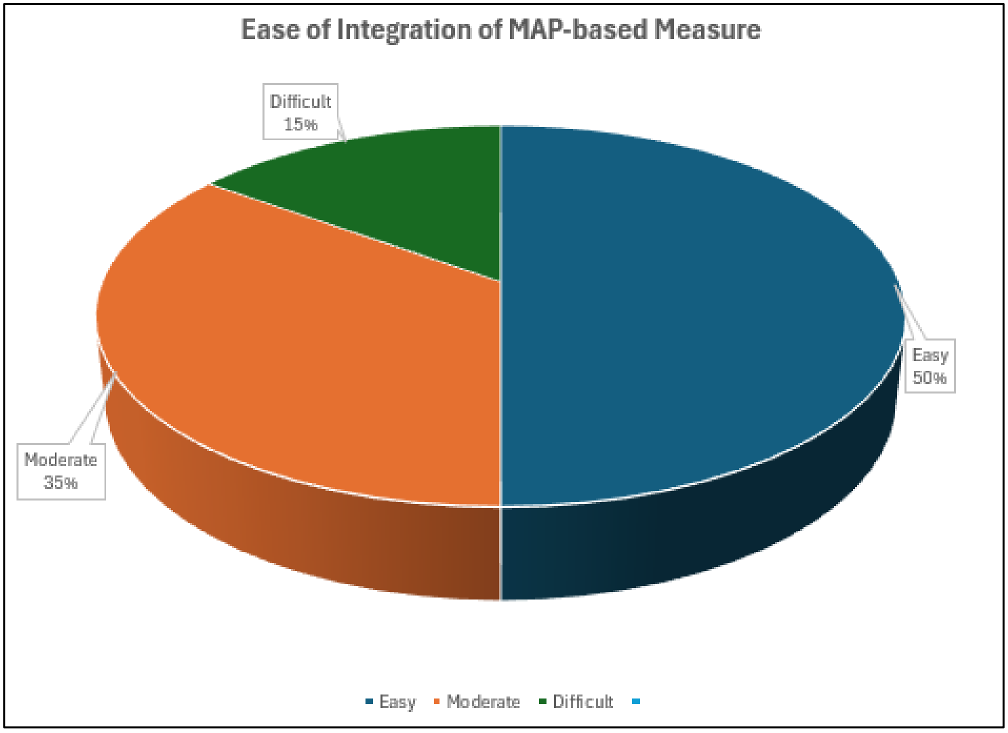

Figure 2: Pie Chart for Ease of Integration of MAP-based Measure.

CONCLUSION

This study demonstrates that the choice of distance metric plays a critical role in classification performance and should be carefully aligned with the underlying characteristics of the dataset. The findings indicate that the Chebyshev distance performs particularly well on datasets with high intra-class variation and irregularities, such as the HODA handwritten digits. Because Chebyshev emphasizes the maximum difference across dimensions, it is more capable of handling the wide variations in stroke shapes, writing styles, and noise commonly found in HODA samples. In contrast, the Euclidean distance shows greater stability and consistency when applied to datasets with more uniform patterns and lower variability, such as the Bangla digits. Its reliance on aggregated squared differences makes it effective when digit shapes are more consistent across samples, which reduces the impact of local distortions or variations.

The incorporation of Mean Average Precision (MAP) as an evaluation metric further enhances the reliability and depth of the classification assessment. Traditional accuracy metrics focus solely on whether the predicted label matches the true label, often overlooking cases where a classifier ranks the correct label close to the top but not first. MAP captures the quality of the entire ranking produced by the classifier, providing a more holistic view of how well each distance metric separates similar and dissimilar instances.

For future research, the study recommends exploring hybrid approaches that integrate traditional distance metrics with modern deep learning techniques. Distance measures could be embedded into feature extraction pipelines, combined with learned embeddings, or used to enhance the interpretability and robustness of neural network–based classifiers. Such integration may leverage the strengths of both worlds distance-based interpretability and deep learning’s representational power to achieve improved accuracy, generalization, and resilience to noise in handwritten digit recognition and other image classification tasks.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM) for the support, facilities, and resources that made this research possible. The authors also extend their special gratitude to the Ingenious Data and Emerging Software Quality Catalysts (IDEAL) Research Group under the Center for Advanced Computing Technology (C-ACT) from UTeM for their continuous guidance, valuable insights, and encouragement throughout the development

of this study.

REFERENCES

1. Yassin, M. M. T. (2023). Comparison on distance metric learning for enhanced precision (Master's thesis). Universiti Teknikal Malaysia Melaka, Melaka, Malaysia.
2. Mulyana, D. I., Hafidz, A., Sumantri, D. B., & Nugroho, K. S. (2022). Identification of Buni fruit image using Euclidean distance method. SinkrOn: Jurnal dan Penelitian Teknik Informatika, 7(2), 392–398. https://doi.org/10.33395/sinkron.v7i2.11333.
3. Suwanda, R., Syahputra, Z., & Zamzami, E. M. Z. (2020). Analysis of Euclidean distance and Manhattan distance in the K-means algorithm for variations in number of centroids K. Journal of Physics: Conference Series, 1566(1), Article 012058. https://doi.org/10.1088/1742-6596/1566/1/012058.
4. Mridha, M. F., Ohi, A. Q., Ali, M. A., Emon, M. I., & Kabir, M. M. (2021). Bangla writing: A multi-purpose offline Bangla handwriting dataset. Data in Brief, 34, 106633. https://doi.org/10.1016/j.dib.2021.106633.
5. Hossein, K., & Ehsanollah, K. (2007). Introducing a very large dataset of handwritten Farsi digits and a study on their varieties. Pattern Recognition Letters, 28(10), 1133–1141. https://doi.org/10.1016/j.patrec.2007.01.002.
6. Arbain, N. A., Azmi, M. S., Ahmad, S. S. S., Muda, A. K., Jalil, I. E. A., & Tiang, K. M. (2017). Dynamic similarity distance with mean average precision tool. Pertanika Journal of Science & Technology, 25(S), 11–18.
7. Prapemrosesan klasifikasi algoritme kNN menggunakan K-means dan matriks jarak untuk dataset hasil studi mahasiswa. (2020). Jurnal Teknologi dan Sistem Komputer, 8(4), 311–316.
8. Alvarez-Melis, D. (2020). Geometric dataset distances via optimal transport. In Advances in Neural Information Processing Systems (Vol. 33, pp. 21428–21439).
9. Kim, J., Cho, S., & Choi, J. (2004). Iris recognition using wavelet features. Journal of VLSI Signal Processing – Systems for Signal, Image, and Video Technology, 38(2), 147–156. https://doi.org/10.1023/B:VLSI.0000040426.72253.b1.
10. Marinov, M., Valova, I., & Kalmukov, Y. (2019, May 16–17). Comparative analysis of existing similarity measures used for content-based image retrieval. In Proceedings of the 2019 X National Conference with International Participation (ELECTRONICA) (pp. 1–4). https://doi.org/10.1109/ELECTRONICA.2019.8825645.
11. Puram, V., Bobbili, R. R., & Thomas, J. P. (2024). Quantum algorithm for Jaccard similarity. arXiv, arXiv:2408.08940.
12. Nurnaningsih, D., Alamsyah, D., Herdiansah, A., & Sinlae, A. A. J. (2021). Identifikasi citra tanaman obat jenis rimpang dengan Euclidean distance berdasarkan ciri bentuk dan tekstur. Building of Informatics, Technology and Science (BITS), 3(3), 171–178. https://doi.org/10.47065/bits.v3i3.1019.
13. Prabiantissa, C. N., Ririd, A. R. T. H., & Asmara, R. A. (2017). Sistem identifikasi batik alami dan batik sintetis berdasarkan karakteristik warna citra dengan metode K-Means clustering. Jurnal Informatika Polinema, 3(2), 26–34. https://doi.org/10.33795/jip.v3i2.10.
14. Hassan, M. R., Hossain, M. M., Bailey, J., & Ramamohanarao, K. (2008, September 15–19). Improving k-nearest neighbour classification with distance functions based on receiver operating characteristics. In Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2008 (LNCS, Vol. 5211, pp. 489–504). Springer. https://doi.org/10.1007/978-3-540-87479-9_50.
15. Arnab, M., Islam, Z., Mamun-Al-Imran, G., & Lasker, E. A. (2021, September 14–16). Iris recognition using wavelet features and various distance-based classification. In Proceedings of the International Conference on Electronics, Communications and Information Technology (ICECIT) (pp. 1–6). Khulna University, Khulna, Bangladesh.
16. Alamri, S. S. A., Bin-Sama, A. S. A., & Bin-Habtoor, A. S. Y. (2016). Satellite image classification by using distance metric. International Journal of Computer Science and Information Security, 14(3), 1–4. https://doi.org/10.6084/m9.figshare.3153877.