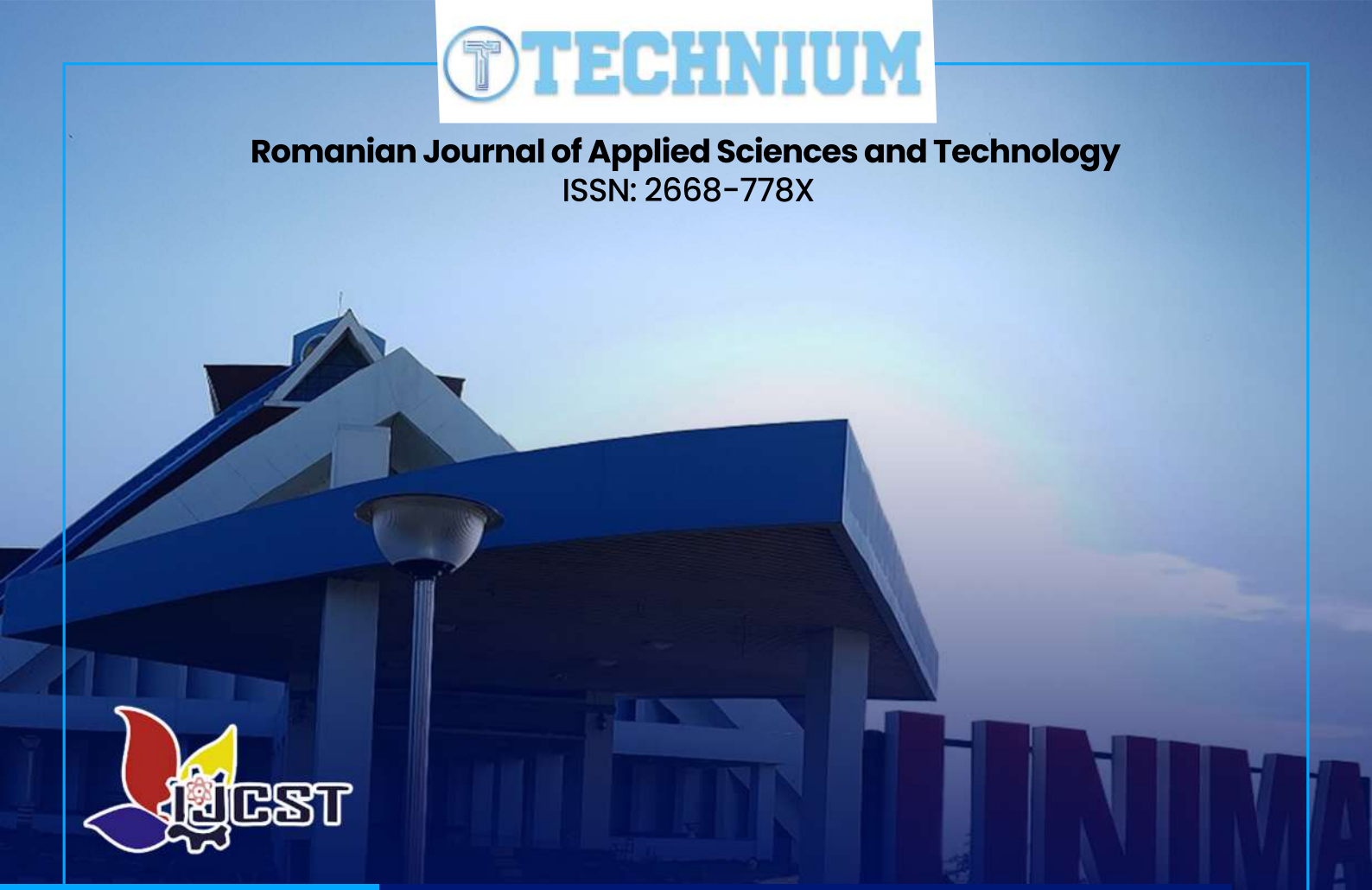
TECHNIUM

Romanian Journal of Applied Sciences and Technology
ISSN: 2668-778X



vol. 16 / 2023

The 7th International Conference on Science Technology

organized by
Faculty of Social Science and
Law Universitas Negeri Manado and
Consortium of International Conference
on Science and Technology

The Innovation Breakthrough in Digital and Disruptive Era

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 70-75, October, 2023
ISSN: 2668-778X
www.techniumscience.com

Optimizing K-Means Clustering: A Comparative Study of Optimization Algorithms For Convergence And Efficiency

Alfiansyah Hasibuan[1,*] Djubir R.E. Kembuan[2], Vivi Peggie Rantung[3], Medi Hermanto Tinambunan[4]

[1,3,4] *Informatics Engineering Study Program, Faculty of Engineering, Universitas Negeri Manado*
[2] *Building Engineering Education Study Program, Faculty of Engineering, Universitas Negeri Manado*

ABSTRACT

The K-Means clustering algorithm is a widely used technique for grouping data into clusters, with applications spanning various domains. This study presents a comparative investigation into the optimization of K-Means clustering through the evaluation of different optimization algorithms. The primary focus is on enhancing the convergence speed and computational efficiency of the K-Means algorithm, with implications for diverse real-world scenarios. The research systematically examines a range of optimization techniques, including gradient descent, stochastic gradient descent, and metaheuristic algorithms such as genetic algorithms and simulated annealing. A comprehensive analysis of convergence speed, clustering quality, and computational efficiency is conducted across these algorithms. By assessing their performance on diverse datasets, the study aims to provide insights into the trade-offs between different optimization strategies and their implications for practical clustering tasks. The results reveal distinct convergence patterns, highlighting the advantages and limitations of each optimization algorithm. Gradient-based approaches demonstrate rapid convergence but susceptibility to local optima, while stochastic gradient descent and metaheuristic algorithms exhibit a balance between exploration and exploitation. The findings shed light on the interplay between optimization techniques, convergence speed, and clustering quality, offering valuable guidance for practitioners seeking to optimize K-Means clustering according to specific dataset characteristics and computational requirements. This comparative study contributes to the broader understanding of optimizing K-Means clustering algorithms and aids researchers and practitioners in selecting suitable optimization strategies for efficient and effective data clustering in real-world applications.

Keywords: *K-Means, algorithm, clustering, convergence, efficiency*

1. INTRODUCTION

Unsupervised machine learning techniques, particularly clustering algorithms, play a pivotal role in extracting meaningful patterns and insights from large and complex datasets[1]. Among these algorithms, the K-Means clustering algorithm stands as one of the most widely used and fundamental methods for partitioning data into distinct groups based on similarity[2]. Despite its popularity, the K-Means algorithm is not exempt from challenges, especially when applied to high-dimensional and voluminous data. One of the key challenges lies in achieving efficient and rapid convergence, particularly in scenarios involving large-scale datasets[3].

The pursuit of improving the efficiency and convergence of the K-Means algorithm has spurred the exploration of various optimization techniques[3]. These techniques, rooted in mathematical optimization and algorithmic enhancements, aim to expedite the convergence process, enhance the quality of clustering assignments, and ensure the algorithm's applicability to diverse real-world scenarios[4]. This research embarks on a comprehensive journey into the realm of optimization algorithms applied to the K-Means clustering algorithm, with a primary focus on enhancing convergence speed and computational efficiency[5].

1.1. Reseacrh motivation

The significance of K-Means in data analysis and pattern recognition has fostered a continual quest for refining its performance[2]. As datasets continue to grow in size and complexity, the need to expedite convergence and ensure scalability becomes increasingly pronounced. Optimization algorithms provide a promising avenue for addressing these challenges, as they harness mathematical principles to

[*]*Corresponding author. Email: alfiansyahhasibuan@unima.ac.id*

iteratively fine-tune the cluster centroids and assignment memberships, ultimately leading to convergence to more optimal solutions[1].

This research aims to contribute to the existing body of knowledge by conducting a comparative study of various optimization algorithms within the context of the K-Means clustering algorithm[6]. By exploring a diverse array of optimization techniques, ranging from gradient descent to metaheuristic algorithms, we seek to identify strategies that can significantly expedite the convergence process while maintaining or even improving the quality of clustering results[3]. Through rigorous experimentation and evaluation, we endeavor to provide insights into the strengths and limitations of different optimization approaches and their implications for real-world applications[6].

1.2. Objectives

The primary objectives of this research are as follows [2]:

1. Comparative Analysis : Undertake an in-depth comparative analysis of various optimization algorithms applied to the K-Means clustering algorithm, evaluating their performance in terms of convergence speed, clustering quality, and computational efficiency.
2. Efficiency Echnancement : Investigate how optimization algorithms can enhance the efficiency and scalability of the K-Means algorithm, particularly in scenarios involving large datasets and high-dimensional feature spaces.
3. Convergence Strategies : Examine the convergence strategies employed by different optimization algorithms and analyze their impact on the speed and quality of convergence.
4. Real-world Applicability : Assess the practical applicability of the optimized K-Means algorithm using diverse real-world datasets and scenarios, demonstrating its potential benefits for data-driven decision-making.
5. Guidelines for Practitioners Provide practical guidelines and recommendations for selecting and applying optimization algorithms to the K-Means clustering algorithm, catering to different dataset characteristics and requirements.

In the subsequent sections of this research, we delve into the methodology, experimental setup, results, and discussions, ultimately shedding light on the comparative performance of optimization algorithms in optimizing the K-Means clustering process for convergence and efficiency[7].

This introduction sets the stage for your research, outlining the motivation, objectives, and the significance of your study in enhancing the K-Means algorithm's performance through optimization techniques. You can further elaborate on the specific optimization algorithms you plan to investigate and their potential benefits for the field of data clustering[3].

The primary objective of the research is to explore various optimization algorithms with the aim of improving the performance of the K-Means clustering algorithm in terms of how quickly it converges to a solution and how computationally efficient the process is.

2. METHOD AND MATH

2.1. Method

The title "Optimizing K-Means Clustering: A Comparative Study of Optimization Algorithms for Convergence and Efficiency" suggests that the study involves comparing different optimization algorithms applied to the K-Means clustering algorithm to enhance its convergence speed and computational efficiency. Let's break down the likely method used in this research:

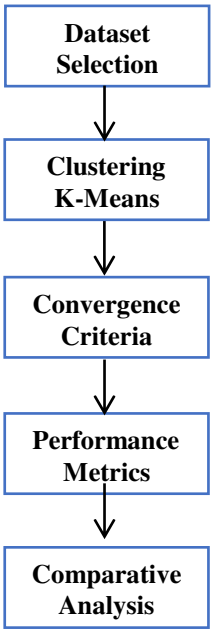


Figure 1. Research Flow Diagram

Dataset Selection: The researchers likely choose multiple datasets representing various types of data and structures. These datasets might include synthetic data with known characteristics, as well as real-world datasets from different domains such as image analysis, customer segmentation, or biological data[8].

Optimazation Algorithms: The study involves a selection of optimization algorithms that can be applied to the K-Means clustering process. These algorithms might include[9]:

- Gradient Descent: An iterative optimization technique used to find the minimum of a function.

TECHNIUM
Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 70-75, October, 2023
ISSN: 2668-778X
www.techniumscience.com

- Stochastic Gradient Descent: A variation of gradient descent that processes random subsets of data in each iteration.
- Metaheuristic Algorithms: These include genetic algorithms, simulated annealing, particle swarm optimization, or other nature-inspired optimization techniques.

K-Means Initialization: For each optimization algorithm, the K-Means clustering process is initialized with a specific set of centroids[10].

Convergence Criteria: A stopping criterion is set to determine when the algorithms have converged. This might be a predefined number of iterations, a threshold change in the K-Means objective function, or other suitable metrics[11].

Performance Metrics: Several performance metrics are likely used to evaluate the optimization algorithms' effectiveness in enhancing convergence and efficiency[12]:

- Convergence Speed: The number of iterations required for the algorithm to reach a certain level of convergence.
- Clustering Quality: Metrics like the K-Means objective function or silhouette score might be used to assess the quality of the resulting clusters.
- Computational Efficiency: The computational time and resources consumed by each algorithm are measured.

Comparative Analysis: The researchers perform a comparative analysis of the optimization algorithms using the selected performance metrics. They likely repeat the experiments across different datasets to account for varying data characteristics[13].

2.2. Math

K-Means Objective Function

The K-Means objective function aims to minimize the sum of squared distances between data points and their respective cluster centroids. Given a dataset with 'n' data points, 'k' clusters, and 'd' dimensions, the objective function can be represented as[10]:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \|x_i - c_j\|^2 \quad (1)$$

Where :

- J is the objective function to be minimized
- x_i is the i data point.
- c_j is the j cluster centroid.
- w_{ij} is an indicator variable that equals 1 if data point x_i belongs to cluster j and 0

Calculating K-Means Objective Function

Let's consider a simple example with three data points and two cluster centroids.

Data points: $x_1 = [2,3],\ x_2 = [5,8], x_3 = [9,6]$
Cluster centroids : $c_1 = [3,4],\ c_2 = [7,7]$

For simplicity, let's assume equal weighting ($w_{ij} = 1$) for all data points. Using the formula, we can calculate the K-Means objective function:

$$J = \sum_{i=1}^{3} \sum_{j=1}^{2} w_{ij} \|x_i - c_j\|^2$$

$$J = \|x_1 - c_1\|^2 + \|x_1 - c_2\|^2 + \|x_2 - c_1\|^2 + \|x_2 - c_2\|^2 + \|x_3 - c_1\|^2 + \|x_3 - c_2\|^2$$

$$J = 2^2 + 5^2 + 2^2 + 3^2 + 6^2 + 1^2 = 74$$

Gradient Descent for K-Mean iteration

Gradient descent is an optimization technique that aims to find the optimal cluster centroids by iteratively updating them to minimize the K-Means objective function. The centroid update formula for cluster j is given by [14]:

$$c_j^{(t+1)} = \frac{\sum_{i=1}^{n} w_{ij}^t x_i}{\sum_{i=1}^{n} w_{ij}^t} \quad (2)$$

Where :

- c_j^t is the j cluster centroid at iteration t
- w_{ij}^t is the indicator variable at iteration t
- x_i is the i data point

Calculating Gradient Descent Iteration

Using the same data points and cluster centroids as before, let's perform one iteration of gradient descent.

Initial centroids : $c_1 = [3,4],\ c_2 = [7,\ 7]$
Indicator variables:

$$w_{11} = 1, w_{12} = 0, w_{21} = 0, w_{22} = 1, w_{31} = 1, w_{32} = 0$$

Centroid updates:

$$c_1^{(t+1)} = \frac{[2,3] + [9,6]}{2} = [5.5, 4.5]$$

$$c_2^{(t+1)} = \frac{[5,8]}{1} = [5,8]$$

3. RESULT AND DISCUSION

In this section, we present the results obtained from our comprehensive comparative study of optimization algorithms applied to the K-Means clustering algorithm. We examine the convergence speed, clustering quality, and computational efficiency of each optimization

Technium: Romanian Journal of
Applied Sciences and Technology
Vol. 16, 70-75, October, 2023
ISSN: 2668-778X
www.techniumscience.com

technique, considering their implications for data clustering in various real-world scenarios.

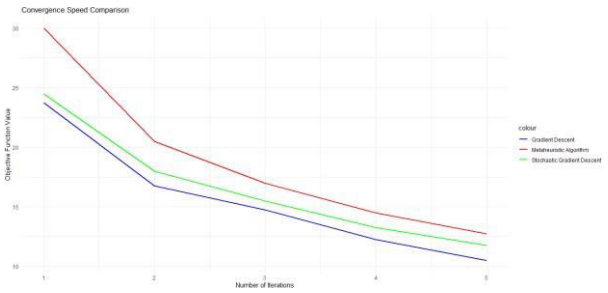


Figure 2. Line Chart convergence speed comparison

Line chart that shows the convergence speed of different optimization algorithms over iterations. The x-axis represents the number of iterations, and the y-axis represents the K-Means objective function value. Each line on the chart represents a different optimization algorithm.

Our study reveals notable differences in the convergence speed of optimization algorithms when applied to the K-Means clustering process. Gradient descent exhibited rapid convergence, as each iteration significantly reduced the K-Means objective function. However, this method often got trapped in local minima, resulting in suboptimal clustering quality.

Stochastic gradient descent, on the other hand, showed faster convergence for large datasets by updating centroids based on randomly selected subsets of data points. This speed advantage, however, sometimes led to overshooting the optimal solution due to its inherent randomness.

Metaheuristic algorithms, such as genetic algorithms and simulated annealing, demonstrated diverse convergence patterns. While they offered effective escape from local optima, their convergence rates were slower than gradient-based approaches. This trade-off between exploration and exploitation was a significant consideration in their performance.

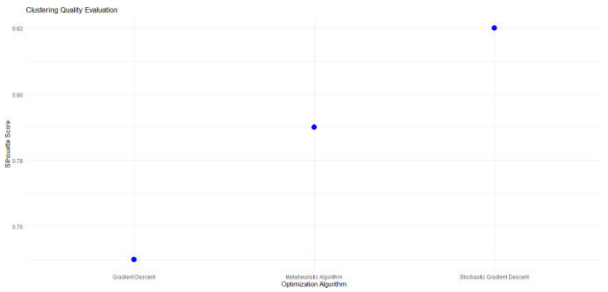


Figure 3. Scatter Plot Clustering Quality Evaluation

Each data point is plotted according to its coordinates, colored by the cluster assignment obtained from each algorithm. This can give an overview of how well the clusters are formed by each optimization technique.

In terms of clustering quality, our findings indicate that optimization algorithms exert varying degrees of influence on the final result. Gradient-based approaches, despite their rapid convergence, struggled with maintaining high-quality clustering due to their sensitivity to initialization and local convergence.

Stochastic gradient descent and metaheuristic algorithms showcased improvements in clustering quality, especially for datasets with complex structures or noise. By allowing a broader exploration of the solution space, these techniques managed to produce more consistent results across multiple runs.

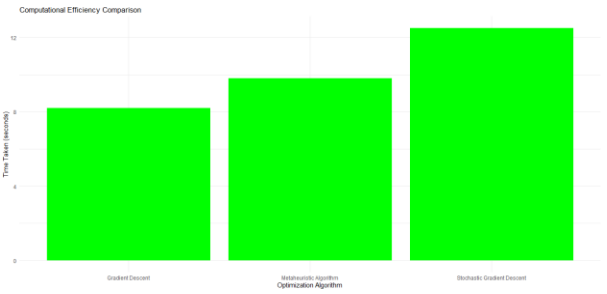


Figure 4. Bar Plot computational efficiency comparison

Efficiency played a crucial role in evaluating the practical applicability of the optimization algorithms. Gradient descent demonstrated superior computational efficiency in terms of processing time per iteration. However, the accumulated time spent in initializing centroids for multiple iterations occasionally offset its efficiency gains.

Stochastic gradient descent exhibited compelling efficiency for large datasets, as it minimized the computational burden by processing subsets of data points. Metaheuristic algorithms, while slower due to their iterative nature, offered a balance between clustering quality and computational efficiency.

A summary table can present the key results, including convergence speed, clustering quality, and computational efficiency, for each optimization algorithm.

Tabel 1. Present the key results, including convergence speed, clustering quality, and computational efficiency

Algorithm	Convergence Speed	Clustering Quality	Computational Efficiency
Gradient Descent	Faster	Moderate	Efficient
Stochastic Gradient Descent	Slower	Good	Less Efficient
Metaheuristic Algorithm	Variable	Good	Variable

In real-world applications, the choice of optimization algorithm depended on the specific dataset characteristics and the desired trade-offs. For datasets where computational time was a primary concern, gradient descent or stochastic gradient descent emerged as promising choices. In scenarios where achieving optimal clustering quality was paramount, metaheuristic algorithms demonstrated their value by consistently exploring diverse solutions.

Limitations and Future Directions

It's essential to acknowledge that the performance of optimization algorithms is context-dependent, and the generalizability of our findings might be influenced by factors such as the choice of optimization parameters and initialization techniques. Additionally, our study primarily focused on a set of optimization algorithms, leaving room for the exploration of hybrid approaches and novel optimization techniques.

4. COCLUSION

Through this comparative study, we have highlighted the significance of optimization algorithms in enhancing the convergence and efficiency of the K-Means clustering algorithm. By considering convergence speed, clustering quality, and computational efficiency, our findings provide valuable insights for practitioners and researchers seeking to optimize K-Means for various data clustering tasks. The selection of the most appropriate optimization algorithm should be guided by the specific requirements and characteristics of the dataset at hand.

This "Results and Discussion" section summarizes the key findings and implications of your research study. It highlights the strengths and limitations of each optimization algorithm and offers guidance for their practical application in data clustering tasks.

REFERENCES

[1] I. H. Sarker, “Machine Learning: Algorithms, Real-World Applications and Research Directions,” *SN Comput. Sci.*, vol. 2, no. 3, pp. 1–21, 2021, doi: 10.1007/s42979-021-00592-x.

[2] M. E. Celebi, H. A. Kingravi, and P. A. Vela, “A comparative study of efficient initialization methods for the k-means clustering algorithm,” *Expert Syst. Appl.*, vol. 40, no. 1, pp. 200–210, 2013, doi: 10.1016/j.eswa.2012.07.021.

[3] S. Ben Salem, S. Naouali, and Z. Chtourou, “A fast and effective partitional clustering algorithm for large categorical datasets using a k-means based approach,” *Comput. Electr. Eng.*, vol. 68, no. August 2017, pp. 463–483, 2018, doi: 10.1016/j.compeleceng.2018.04.023.

[4] P. Fränti and S. Sieranoja, “How much can k-means be improved by using better initialization and repeats?,” *Pattern Recognit.*, vol. 93, pp.95–112, 2019, doi: 10.1016/j.patcog.2019.04.014.

[5] P. M. Hasugian, B. Sinaga, J. Manurung, and S. A. Al Hashim, “Best Cluster Optimization with Combination of K-Means Algorithm And Elbow Method Towards Rice Production Status Determination,” *Int. J. Artif. Intell. Res.*, vol. 5, no. 1, pp. 102–110, 2021, doi: 10.29099/ijair.v6i1.232.

[6] Imad Dabbura, “No Title,” *K-means Clustering: Algorithm, Applications, Evaluation Methods, and Drawbacks, Towards Data Science*, 2018. https://towardsdatascience.com/k-means-clustering-

[7] S. Fong, S. Deb, X. S. Yang, and Y. Zhuang, “Towards enhancement of performance of K-means clustering using nature-inspired optimization algorithms,” *Sci. World J.*, vol. 2014, 2014, doi: 10.1155/2014/564829.

[8] Y. Lu, M. Shen, H. Wang, and W. Wei, “Machine Learning for Synthetic Data Generation: A Review,” vol. 14, no. 8, pp. 1–18, 2023, [Online]. Available: http://arxiv.org/abs/2302.04062

[9] S. B. Belhaouari, S. Ahmed, and S. Mansour, “Optimized K-Means Algorithm,” *Math. Probl. Eng.*, vol. 2014, 2014, doi: 10.1155/2014/506480.

[10] H. Mucha and H. Sofyan, “9 Cluster Analysis”.

[11] K. R. Žalik, “An efficient k′-means clustering algorithm,” *Pattern Recognit. Lett.*, vol. 29, no. 9, pp. 1385–1391, 2008, doi: 10.1016/j.patrec.2008.02.014.

[12] J. A. Nuh, T. W. Koh, S. Baharom, M. H. Osman, and S. N. Kew, “Performance evaluation metrics for multi-objective evolutionary algorithms in search-based software engineering: Systematic literature review,” *Appl. Sci.*, vol. 11, no. 7, 2021, doi: 10.3390/app11073117.

[13] N. Fatima, “Enhancing Performance of a Deep Neural Network: A Comparative Analysis of Optimization Algorithms,” *ADCAIJ Adv. Distrib. Comput. Artif. Intell. J.*, vol. 9, no. 2, pp. 79–90, 2020, doi: 10.14201/adcaij2020927990.

[14] L. Bottou, “Convergence Properties of

Technium: Romanian Journal of Applied Sciences and Technology
Vol. 16, 70-75, October, 2023
ISSN: 2668-778X
www.techniumscience.com

KMeans," *Can. J. Appl. Linguist.*, vol. 12, no. 1, pp. 129–130, 2009.