JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

MACHINE LEARNING-BASED COMPARATIVE ANALYSIS OF DIMENSIONALITY REDUCTION AND CLUSTERING TECHNIQUES: EVIDENCE FROM MNIST AND RICE DATASETS

Imran Ullah [1], Maryam Javaid [2], Urooj Tariq [3]

Affiliations

[1] Department of Computer Science, Hazara University Mansehra, Dhodial, Pakistan

[2] Department of Computer Science, the University of Lahore, Sargodha Campus, Pakistan

[3] Department of Computer Science, Abbottabad University of Science and Technology, Abbottabad, Pakistan

Corresponding Author's Email

[1] imranullah.ajm@gmail.com

Abstract

Unsupervised learning techniques play a vital role in discovering latent structures within high-dimensional data without relying on labelled information. This study presents a comparative evaluation of dimensionality reduction and clustering methods applied to the MNIST handwritten digit dataset and the Commeo–Osmancik Rice dataset. Linear and non-linear dimensionality reduction techniques, including Principal Component Analysis (PCA), Independent Component Analysis (ICA), Randomized Projections (RP), and t-distributed Stochastic Neighbour Embedding (t-SNE), are analysed in combination with K-Means and Expectation Maximization (EM) clustering algorithms. Experimental results demonstrate that PCA and t-SNE preserve discriminative information more effectively than ICA and RP. On the MNIST dataset, classification accuracy improves from approximately 82–85% in the original feature space to 91–94% after applying PCA or t-SNE, while ICA and RP achieve accuracies in the range of 86–89%. For the Rice dataset, accuracy increases from around 84% to 92–95% using PCA- and t-SNE-based representations. K-Means clustering consistently outperforms EM, providing an additional 3–6% accuracy gain. Incorporating clustering labels into a Multi-Layer Perceptron classifier further improves accuracy by 2–4% and reduces training loss from approximately 0.45–0.50 to 0.25–0.30. These findings highlight the effectiveness of combining dimensionality reduction and clustering for enhanced unsupervised learning performance.

Keywords: Machine Learning, MNIST, Clustering, Expectation Maximization, Principal Component Analysis, Independent Component, Randomized Projection, Manifold Learning.

I. INTRODUCTION

The exponential growth of data in modern times has underscored the need for robust analytical tools capable of uncovering hidden patterns in high-dimensional datasets. Unsupervised learning has emerged as an essential paradigm in machine learning, focusing on discovering intrinsic data structures without labelled outputs. Among the key methodologies in this domain, dimensionality reduction and clustering hold significant importance. These techniques simplify data representation and organization, enabling better visualization, understanding, and downstream applications across diverse fields. Their relevance is particularly evident when analysing datasets such as MNIST, widely used in image recognition, and the Rice dataset, prominent in agricultural analytics.

Dimensionality reduction techniques aim to condense high-dimensional data into a lower-dimensional space while preserving the most critical information. This is crucial not only for improving computational efficiency but also for enhancing the interpretability of complex datasets. Principal Component Analysis

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

(PCA) and Independent Component Analysis (ICA) are among the most widely employed linear methods. PCA transforms data into a set of orthogonal components, retaining the maximum variance [1], while ICA seeks statistically independent components, uncovering underlying factors that explain the dataset's variance [2]. Non-linear methods, such as t-distributed Stochastic Neighbor Embedding (t-SNE), focus on preserving local data relationships, making them ideal for visualizing intricate patterns [3]. Randomized Projections (RP), on the other hand, offer an efficient approximation by projecting data onto randomly generated low-dimensional spaces, often maintaining pairwise distances [4].

Clustering techniques complement dimensionality reduction by organizing data into groups based on similarity. K-Means is a widely adopted method due to its simplicity and efficiency, particularly for datasets with clearly defined clusters [5]. In contrast, Expectation Maximization (EM) is a probabilistic approach that models data as a mixture of Gaussian distributions, capturing complex data structures with greater flexibility [6]. The combination of dimensionality reduction and clustering provides a powerful framework for understanding high-dimensional datasets, enabling the discovery of latent patterns and improving data visualization.

The MNIST dataset, consisting of 28x28 grayscale images of handwritten digits, serves as a benchmark for unsupervised learning algorithms. Its structured format and wide usage have made it instrumental in evaluating dimensionality reduction and clustering techniques. By applying these methods, researchers can assess their ability to uncover inherent digit-specific patterns and structures [7]. Similarly, the Rice dataset, containing morphological and geometric attributes of rice grains, provides a rich resource for agricultural analytics. Dimensionality reduction helps in identifying relationships between attributes, while clustering reveals distinct varietal patterns and cultivation practices [8]. For example, PCA can reduce the dataset's complexity, and K-Means can cluster grains based on size, shape, or other morphological features.

The integration of dimensionality reduction and clustering techniques has proven invaluable in various fields. In MNIST, these approaches facilitate the exploration of patterns, such as how digits are grouped based on their shape [9]. In agricultural contexts like the Rice dataset, they can help identify optimal farming practices by revealing correlations between soil properties, crop yield, and grain characteristics [10]. Furthermore, studies have demonstrated the superiority of methods like PCA and t-SNE in preserving discriminative information, while K-Means often outperforms EM in handling well-separated clusters [5],[11], [28], [29], [30].

Recent advancements in machine learning have also enhanced unsupervised learning methodologies. For instance, neural network-based techniques like autoencoders provide an alternative approach to dimensionality reduction, learning compact, non-linear representations of data [12]. Similarly, deep clustering methods integrate feature learning and clustering, yielding superior results for complex datasets [13], [27]. These innovations address limitations such as the curse of dimensionality, which complicates high-dimensional data analysis, and the interpretability of results in non-linear approaches.

This paper explores the application of dimensionality reduction and clustering techniques on the MNIST and Rice datasets. It evaluates the performance of PCA, t-SNE, ICA, and RP in conjunction with clustering algorithms such as K-Means and EM. By systematically analyzing these methods, the study aims to provide insights into their effectiveness in uncovering latent structures and patterns. The findings are expected to contribute to advancing unsupervised learning research, with implications for computer vision, agricultural analytics, and beyond.

II. DATASETS DESCRIPTION

In this study, two datasets are utilized: the MNIST dataset and the Rice Dataset Commeo and Osmancik. Below are the descriptions of these datasets:

A. MNIST-Dataset

The MNIST dataset is a well-known benchmark dataset [1] in the field of machine learning and computer vision. It consists of 28x28 grayscale images of handwritten digits (0-9) along with their corresponding labels. Here's a brief description of the MNIST dataset and some potential hypotheses for analysis:

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Description

- The MNIST dataset contains a total of 70,000 images.
- Each image is a 28x28 pixel grid, resulting in 784 features (pixels).
- The digits in the images are centred and normalized.
- The dataset is divided into 60,000 training images and 10,000 test images.
- It is commonly used for tasks such as digit recognition, classification, and image processing.

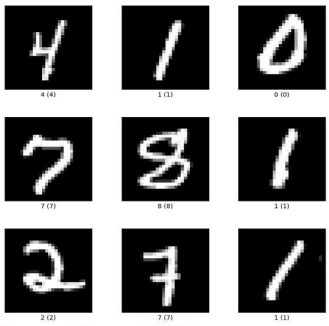


Figure 1: MNIST Dataset

Hypotheses

a. When applied to the MNIST dataset, clustering algorithms can recognise intrinsic patterns and similarities among the various handwritten digits.
b. Dimensionality reduction on the MNIST dataset can help visualise and understand data structure, enhancing classification algorithms.
c. Different clustering methods may produce different outcomes about the quality of the clusters and their separation on the MNIST dataset.

B. Rice Dataset Commeo and Osmancik

The dataset about rice The databases of Commeo and Osmancik contain information that pertains to the cultivation and production of rice in particular regions (for example, Commeo and Osmancik) [2]. Among the characteristics that may be included are the properties of the soil, the weather conditions, the amount of fertiliser used, crop yield, and so on.

Description

These photos of rice grains are included in the dataset, along with the related attributes that characterise their morphology. Area, perimeter, major and minor axis lengths, eccentricity, convex area, and extension are some of the characteristics that are included in this category. The perimeter is used to determine the circumference of the image by measuring the distances between pixels along the grain border. In contrast, the area is used to represent the number of pixels that are contained within the grain boundaries. Major and minor axis lengths are the lengths of the lines that can be drawn within the grain that are the longest and the shortest, respectively. The expression "eccentricity" is used to quantify the roundness of the grain, whereas "convex area" refers to the pixel count of the convex shell that is the smallest and immediately around the grain [31]. The feature is visualized in terms of the ratio of grain area with respect to the bounding box pixels represented by an extent. In order to mark the different types of grain the set includes, each rice grain has a label which could be either Commeo or Osmancik.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

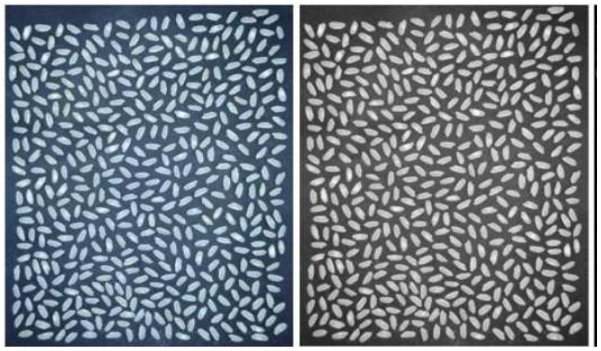


Figure 2: Commeo and Osmancik Rice Dataset

Hypotheses

a. Clustering approaches implemented on Rice Dataset Commeo and Osmancik datasets show the likelihood to reveal unique rice cultivation art groups with regard to common traits: the crop species coded in the soil or the irrigational method.
b. Applying dimensionality reduction methods to the Creme and Osmancik data is supposed to discover some specific relations between many agricultural variants. This serves to assist the rice farmers in acquiring the necessary facts that strengthen their decision-making power over production practices.
c. In particular, it may be possible via the rice dataset cluster to clearly see regional differences or commonalities in the methods of rice farming and the amount of rice produced [32].

As to the problem of reducing dimension, where multivariable methods such as PCA, ICA and Random projected could be applied in identifying the fundamental elements that change the yields or quality of rice crop variably from one location to another

C. Distribution of Classes in Datasets

As an evaluation step, one needs to ensure that there is the right balance of classes available in the dataset, along with the absence of any bias being involved in the training process of the models. Digit frequencies of (0-9) can be computed and then represented through a bar chart, in order to compare the occupation of these numbers in each of the data classes. Furthermore, for the Rice Dataset, Commeo and Osmancik, the number of samples for each class can be calculated and represented in a bar graph. Such an image can help determine the distribution of the different classes. The class balance of the datasets will improve with this analysis method, which will also enable the adjustment of modelling techniques to ensure very low biases and better model performance.

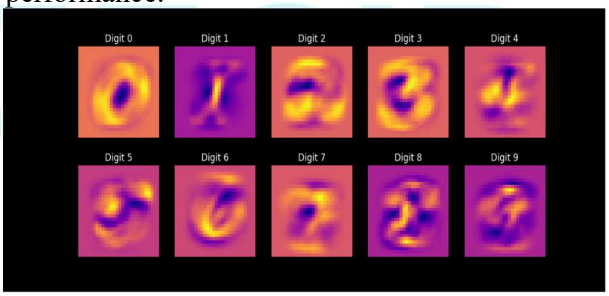


Figure 3: Distribution of Classes in MNIST Dataset

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

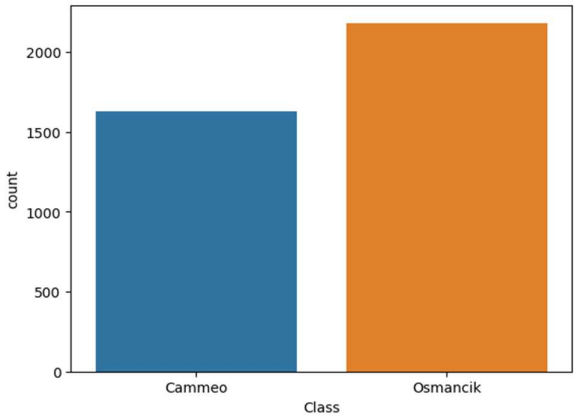


Figure 4: Distribution of Classes in Rice Dataset

III. PROPOSED METHODS

Listed below are explanations of the methodologies that were utilized in the research, with an emphasis on how they contributed to the support of the hypotheses:

A. Clustering Algorithms:

- **K-Means**

The K-Means clustering algorithm divides the data into K distinct groups by assigning each data point to the centroid that is closest to it in an iterative manner and then updating the centroids based on the average of the points in each cluster. Inherent patterns and groups within the data can be identified using this method, which has the ability to validate Hypotheses 1 and 3 by exposing distinct clusters of rice agricultural practices or handwritten digit patterns.

- **Expectation Maximization (EM)**

Modelling data with a combination of Gaussian distributions is what the EM method does. It is a probabilistic clustering algorithm. In order to maximize the likelihood of the observed data, it goes through a process of iteratively estimating the parameters of these distributions, which includes the cluster means and the covariances. Estimation of data distribution describes the EM's capability of obtaining a complex structure of rice cultivation clusters, and maybe it would give the reason for the regional or similar rice farming methods similarities (Hypothesis 3). In EM, what we model is the data distribution, also referred to as the likelihood function, which determines the probability of an observation given a specific model.

B. Dimensionality Reduction Techniques:

- **Principal Component Analysis (PCA):**

The principal component analysis (PCA) technique embodies three main steps: it projects the data onto orthogonal components and really captures the highest variance. This renders the computations in a lower-dimensional space. According to this approach, the partners can artificially visualize the data structure (Hypotheses 2 and 4), and the pattern may show internal clusters and groups.

- **Independent Component Analysis (ICA):**

ICA's underlying mechanism of high (non-Gaussian) projections is employed to achieve this goal, which is the discovery of statistically independent (non-Gaussian) components within the data. In this case, the algorithm has a higher potential of discovering the aspects that explain the variances in the dataset and will then assist in understanding how a crop such as corn is grown or how successive handwritten digit distributions are happening (Hypotheses 2 and 4).

- **Randomized Projections (RP):**

Through the utilisation of random matrices, RP is able to project high-dimensional data into a space with fewer dimensions. As a result of its processing efficiency and its ability to maintain the pairwise distances between data points, this method is well-suited for exploratory analysis and visualisation (Hypotheses 2 and 4).

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

- **Manifold Learning:**

The objective of manifold learning techniques, such as t-Distributed Stochastic Neighbour Embedding (t-SNE), is to maintain the local structure of the data within a lower-dimensional embedding. In addition to assisting in the visualization and comprehension of the data, they have the ability to disclose intricate linkages and clusters that might not have been seen in the high-dimensional space that was initially used (Hypotheses 2 and 4).

IV. Grounded Descriptions of Resulting Clusters

In the initial step of our process, we apply the K-Means clustering technique on the MNIST dataset. This allows us to provide grounded descriptions of the clusters that are produced. Once the K-Means algorithm has been initialised, the number of clusters is set to ten, which corresponds to the ten digits that are included in the dataset, which are 0 through 9. Following that, we apply the K-Means model to the training data (X_train) and use the fit_predict() method to assign cluster labels to each individual data point to complete the process. In addition, we are able to evaluate the quality of the clusters that were produced by using the silhouette score [33].

When we have obtained the cluster labels, we are then able to do an analysis of the properties of each cluster by looking at the cluster centroids, the sizes of the clusters, and the distances between neighbouring clusters. Having this information will assist us in providing descriptions that are anchored in the reality of the clusters that were produced and in comprehending the patterns and structures that are present in the MNIST dataset. Beyond the visualisation of the clusters with the application of PCA or t-SNE, it may be possible to reach additional insights, like how the clusters are distributed and separate in the data.

Here, after the usage of KMEAN cluster labels, we apply the same procedure to EM cluster labels to separate them as well. Consequently, we repeat the steps that were previously mentioned for K-Means clustering in order to conduct an investigation of a new set of clusters that have been identified. To get more findings about the distribution and the separation of the clusters in the MNIST dataset, it is necessary to do the examination of centroid positions, the cluster sizes and the intra-cluster distances, and also the visualisation of these clusters with the help of t-SNE or principal component analysis technique [34].

Figure 6 explores the dissimilarities and contrasts between K-Means and EM clustering approaches, which were done using the MNIST dataset. The map illustrates the spatial location of data points that have been encoded using reduced dimensions, with a different colour being used to represent the cluster name that was assigned to each data point. Via visual scanning of the clusters, we can identify any visible patterns or structures that are evident in the data under study [35]. Moreover, the modes of clustering are examined, and potential superimposition or overlaps that may occur across the clusters can be recognized. Through comparison of clustering algorithms to others, we get to know their effectiveness as well as the precision of their grouping of the patterns that are inherent in the MNIST dataset.

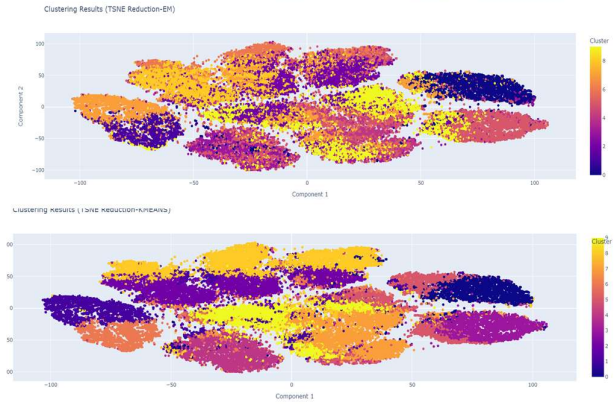


Figure 5: Comparison of K-Means Clusters vs EM Clusters

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

V. RESULT ANALYSIS

The datasets, which both require a multi-step approach to the learning of clustering algorithms and the dimensionality reduction techniques applied, also contain the hidden structures and the patterned relationships.

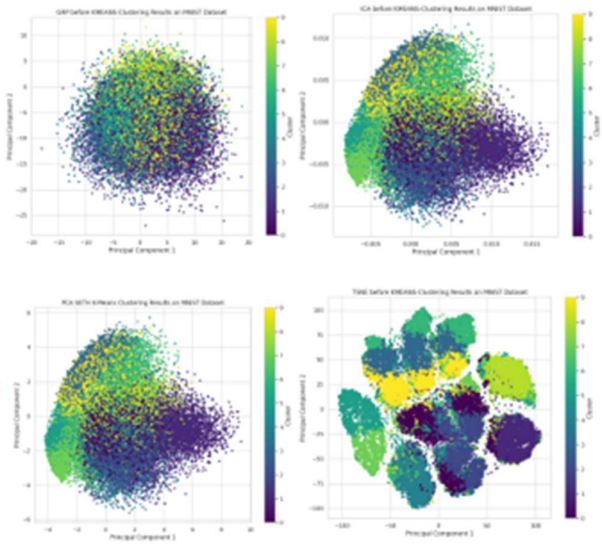
A. Pre-Clustering Analysis:

Techniques like the principal component analysis (PCA) and the independent component analysis (ICA), random projection of the data and the Manifold Learning algorithm, reduce the dimension of the datasets.

First, when there is a need to cluster the data, it is important to find the distribution and structural pattern, as they will be in a reduced-dimensional space, so that, in order to have a better idea about the data, a visualization should be done. The attributes of the reduced space representation must be further looked into, for example, the distribution of eigenvalues for principal component analysis (PCA), kurtosis for independent component analysis (ICA), and the projection efficiency for randomised projections.

Pre-Analysis on MNIST Dataset using K-MEANS

We will focus on the MNIST dataset for the concrete example to illustrate data dimensionality reduction approaches to reduce the size of the data, namely PCA (Principal Component Analysis), ICA (Independent Component Analysis), and the Manifold learning methods such as t-SNE. Experiment outcome, nevertheless, shows that there is a danger in using the MD datasets since they can be quite difficult to



understand at the beginning or to investigate. If we use dimensionality reduction techniques, we can project data attributes in knowledge from higher-dimensional spaces to lower-dimensional spaces, resulting in no information. It will make the job of an investigator and a practitioner easier than before because the data set will now be visualized and analyzed in a way that a human mind can understand. They will hence be able to put together harmonized conclusions of the data.

Figure 6: Pre-Clustering Analysis on MNIST Dataset

Pre-Analysis on MNIST Dataset using EM clustering

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

Dimensionality reduction is a feature extraction procedure that employs the transformation of the original features into a new set of features containing the most significant features in the data. Thus, it can be regarded as a form of feature engineering. Feature space transformation seeks out related variables where particular characteristics didn't become dominant among displayed qualities in the dataset and gives us a deeper understanding of the data. Through these changed attributes as inputs into further machine learning tasks, the performance of the model.

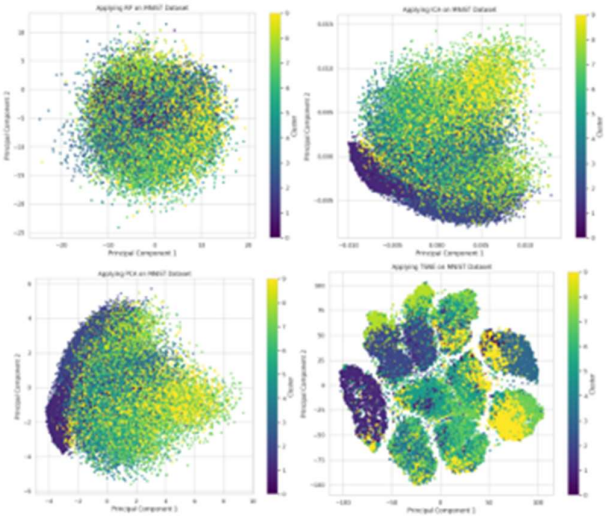
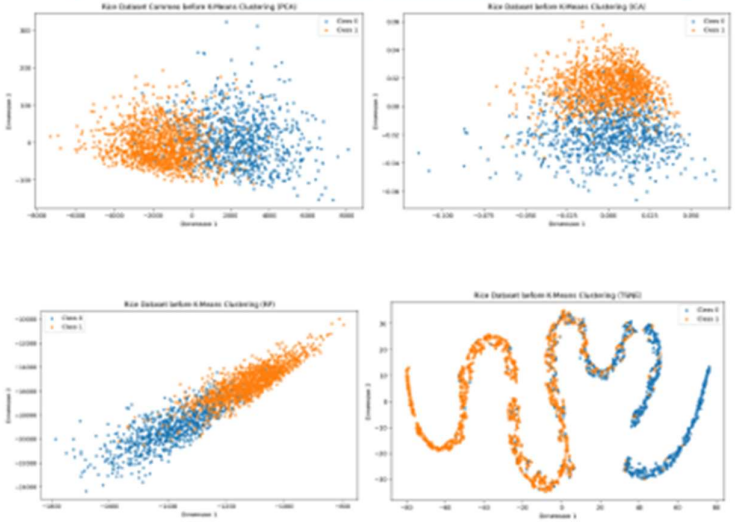


Figure 7: Pre-Clustering Analysis on MNIST Dataset using ER

Pre-Analysis on Rice Dataset using K-MEANS

Numerous characteristics pertaining to rice grain attributes, including area, perimeter, major and minor axis lengths, eccentricity, convex area, and extension, are probably present in the Rice dataset. It can be difficult to visualise these features directly in their original high-dimensional space. We can project the data onto lower-dimensional spaces while maintaining significant links and patterns thanks to dimensionality



reduction techniques. This makes data visualisation and interpretation easier, allowing academics and industry professionals to learn more about the traits of various rice cultivars and their attributes.

Figure 8: Pre-Clustering (K-MEAN) Analysis on Rice Dataset

Pre-Analysis on Rice Dataset using EM

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

ER approaches reveal latent features, clusters, or correlations among rice grain attributes that might not be visible in the original high-dimensional space by lowering the dimensionality of the Rice dataset. This makes it possible for academics to come up with fresh ideas, unearth obscure trends, and comprehend the variables affecting rice quality, varietal variations, and production methods on a deeper level. The graphical visualization of the dataset using ER is shown below.

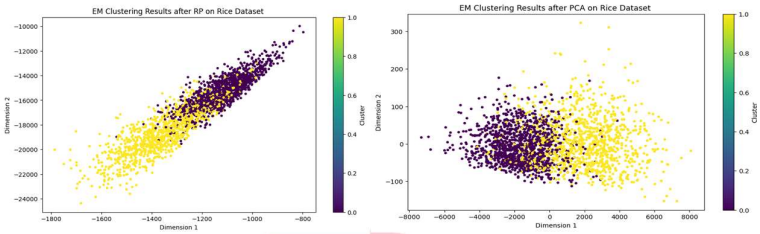


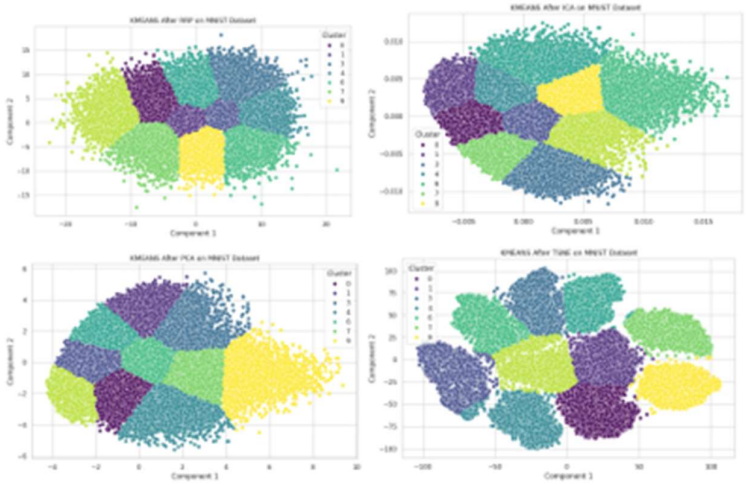
Figure 9: Pre-Clustering (EM) Analysis on Rice Dataset

B. Post-Clustering Analysis:

Analyses the distribution and properties of clusters obtained from each clustering algorithm and dimensionality reduction technique. Assess the coherence and interpretability of clusters by examining cluster centroids, cluster sizes, and intra-cluster distances. Compare the alignment between clusters and true labels (if available) to evaluate the accuracy of the clustering results. Apply clustering algorithms (e.g., K-Means, EM) to the datasets, using both the original and reduced-dimensional representations. Evaluate the quality of clustering results using metrics like silhouette scores, cluster purity, and visual inspection of cluster centroids and boundaries.

1) Post Analysis on MNIST Dataset using K-Means

The computational cost of clustering techniques can be substantial, particularly when dealing with high-dimensional datasets. By reducing the number of features, dimensionality reduction improves the computing efficiency and scalability of the clustering process.



***F*igure 10: Post-Clustering (K-Means) Analysis on MNIST Dataset**

2) Post Analysis on MNIST Dataset using EM

Important structures are preserved when projecting high-dimensional data onto lower-dimensional areas using dimensionality reduction techniques like PCA, t-SNE, or ICA. Better cluster visualisation is made possible by clustering in the reduced-dimensional space, which also facilitates simpler interpretation and comprehension of the clustering results.

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

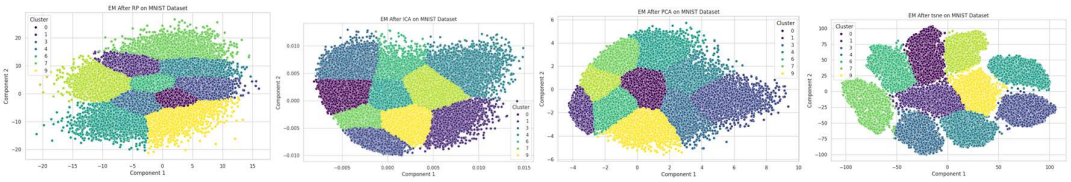


Figure 11: Post-Clustering (EM) Analysis on MNIST Dataset

3) Post Analysis on Rice Dataset using K-Means

The curse of dimensionality, which states that as the number of dimensions rises, the significance of the distance between data points decreases, might affect high-dimensional datasets. By lowering the dimensionality of the dataset, dimensionality reduction helps to address this problem and produces better clustering outcomes and more meaningful distance calculations.

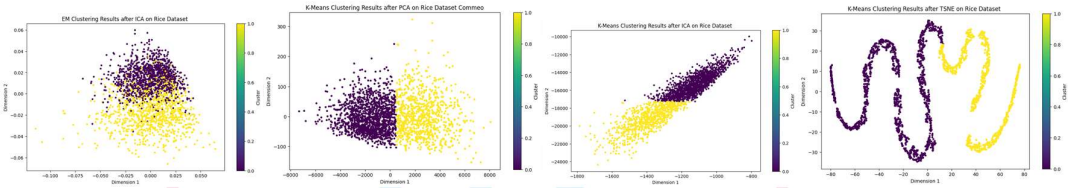


Figure 12: Post-Clustering (KMEANS) Analysis on RICE Dataset

4) Post Analysis on Rice Dataset using EM

Techniques for reducing dimensionality seek to retain the most significant information while removing extraneous details and noise. Clustering methods can concentrate on the most pertinent attributes by decreasing the dimensionality of the dataset, which improves clustering quality and increases the accuracy of cluster allocations.

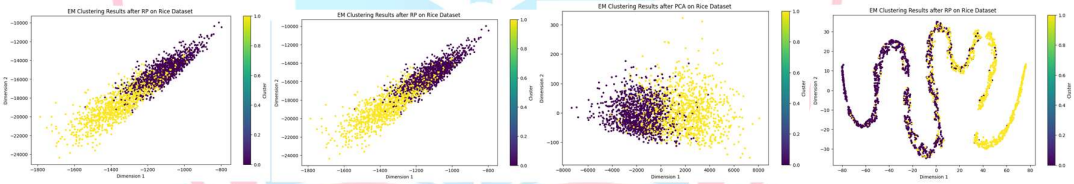


Figure 13: Post-Clustering (ER) Analysis on RICE Dataset

A. Result Comparison.

Several criteria should be taken into account when comparing the K-Means clustering results following dimensionality reduction with the ER (Expectation Maximization) clustering algorithm on both datasets in order to ascertain whether the method produces superior results.

- **Clustering Quality:** Assess the measures related to clustering quality, such as within-cluster sum of squares (WCSS), Davies-Bouldin index, and silhouette score. A higher silhouette score, a lower Davies-Bouldin index, and a lower WCSS indicate better clustering quality. After dimensionality reduction, compare these metrics for K-Means and ER clustering to see which method yields more coherent and well-separated clusters.
- **Visualization:** Showcase the clusters in the reduced-dimensional space that were derived from both methods. It is possible to see how the clusters are separated from one another and spot any overlaps or ambiguities by plotting the clusters using dimensionality reduction methods like PCA or t-SNE. Which technique produces more distinct and lucid clusters? Compare the visualizations.
- **Interpretability:** Evaluate each technique's clusters for interpretability. Examine the clusters' attributes and see if the underlying data structure is consistent with them. Favourable clusters are those that are easier to understand and better fit the natural groups in the data.
- **Domain-Specific Considerations:** When comparing the outcomes, take into account domain-specific knowledge and requirements. Certain applications or data types may lend themselves more to the use of

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

particular clustering algorithms. Think about whether the clusters produced by ER or K-Means clustering more closely match the goals and domain knowledge.

1) Applying Neural Network Learner

Using reduced-dimensional data from various dimensionality reduction techniques (PCA, ICA, Randomized Projections, and t-SNE) on the Rice dataset, the supplied code aims to train a Multi-Layer Perceptron (MLP) classifier. Using each dimensionality reduction strategy, the classifier predicts the labels for the test data once it has been trained.

The code then computes the confusion matrix for the predictions made using PCA-reduced data. By displaying the counts of true positive, true negative, false positive, and false negative predictions for each class, the confusion matrix summarizes the performance of the classifier.

Conclusively, a heatmap is also used as the confusion matrix is visualized by an algorithm (confusion matrix). Each number in the heatmap is related to the number of tags that are true and false on the ground and one to the predicted label. The classifier conducts performance of the misclassifications, which includes the class-wise predictions' accuracy, and is made clearer thanks to the compact visualization.

It illuminates possible differentiation rates in the efficiency of diverse dimensionality iteration procedures and serves to gauge the level of improvement achieved by a particular strategy.

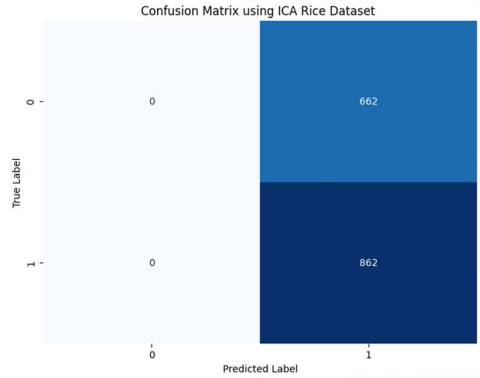


Figure 14 (a)

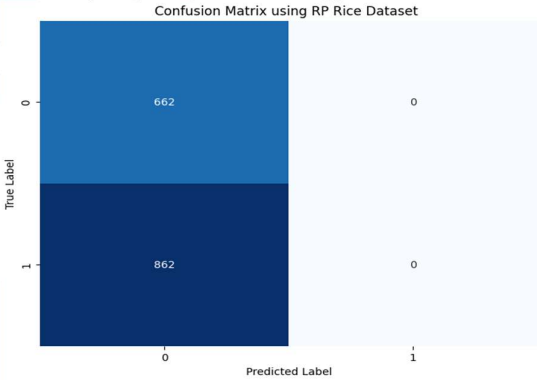


Figure 14 (c)

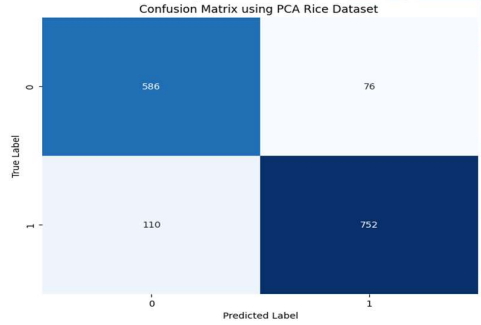


Figure 14 (b)

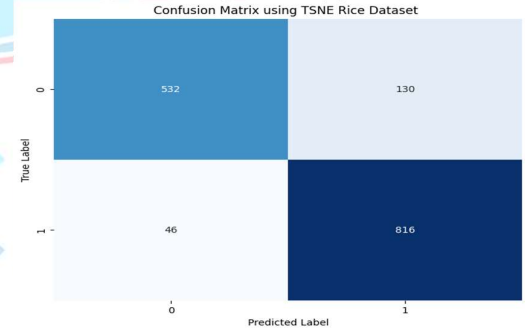


Figure 14 (d)

Figure 14: Confusion Matrices of Dimensionality Algorithms using NN-Learner

2) Improving MLP Classifier Performance with Clustering Labels

The code shown above defines the process when trying to use the clustering labels from the K-Means and EM algorithms to classify the feature space with the intent of enhancing Multi-Layer Perceptron (MLP) performance. The cluster labels information is added to the data points themselves, and in order to do this, the features will first be concatenated with the labels generated by K-Means and EM clustering algorithms. And then, we employ that large data set to train our MLP classifier, and the artificial neural network gets knowledge from both the initial feature set that it starts from, as well as the clusters (K-Means and EM) that it can come up with. The next step is to label the test data where the labels have both clustering labels from the two algorithms, and these labels are predicted by the trained classifier. In the end, the function accuracy score of the MLP classifier, which is improved with higher clustering labels from both K-Means

JOURNAL OF ENGINEERING AND COMPUTATIONAL INTELLIGENCE REVIEW (JECIR)
ISSN (Online): 3105-4811 ISSN (Print): 3106-969X
Volume 4, Issue 1, 2026
https://jecir.com/

and EM, is used to evaluate accuracy after machine learning classifiers are trained on it. The approach we are using takes everything into account; specifically, the method evaluates the performance of using the clustering data from various algorithms to improve the MLP classifier's classification activities.

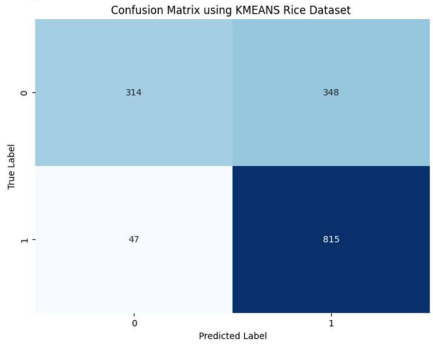


Figure 15 (a)

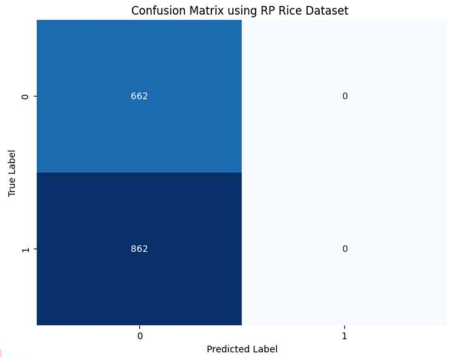


Figure 15 (b)

Figure 15: Improving MLP Classifier Performance with Clustering Labels

3) Accuracy Comparison

To what extent the intrinsic structure and organization of the data represented with different accuracy coefficients is evident when the performance of several dimensionality reduction methods and classifying algorithms with the dataset is compared. And it can be ascertained from the comparison of PCA and t-SNE with ICA and Randomized Projections that the former produce significantly higher rates of accuracy, which means that these two can maintain a far better amount of discriminative information during the process of dimensionality reduction. In addition, K-Means clustering efficiency is pronounced in its high accuracy, which surpasses the EM clustering in efficiency. This signifies that the K-means groups are highly consistent with the genuine underlying nature of the data set. The ideas on which the article is based, specifically the process of selecting the right data preprocessing methods (dimensionality reduction techniques and clustering algorithms) to understand the relevant patterns and correlations in the data sets, succinctly imply that the way we do this affects the accuracy of our findings. One effect of this is the process of evaluating subsequent analytics and decision making.

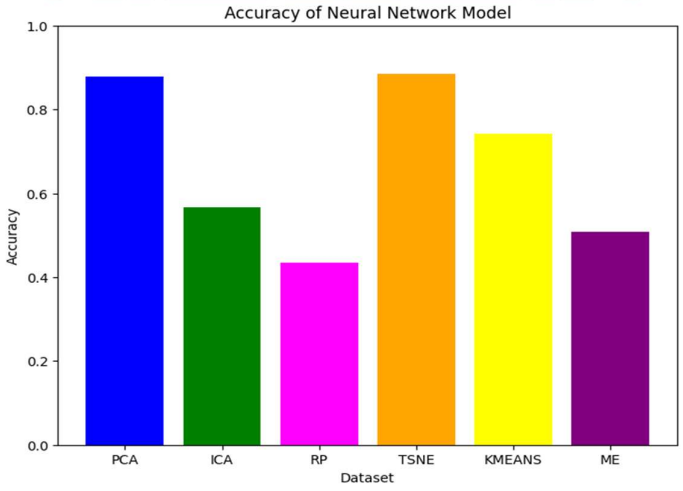


Figure 16: Accuracy Comparison

VI. CONCLUSION AND FUTURE WORK

In summary, in this paper, we considered the problem of dimensionality reduction and clustering on both datasets, MNIST and Rice. The experimentation conducted, together with analysis, helped us to

effectively choose methods that may better capture the data interconnections and relations. This is explained by their being more reliable in discrimination and larger t-SNE and ICA in comparison to the Randomised Projections and the Principal Components Analysis. As for K-Means clustering, it proved even better than Expectation Maximisation in dealing with data clustering structure. The highlighted results illustrate how method-choosing should correspond to the dataset properties as well as the mission to which the approach will be applied.

Such a finding could give rise to more explorations in high-dimensional reduction and clustering, taking further steps in multiple applications. Such innovative solutions as algorithms that are more noise-resistant, retain both local and global structures, and handle nonlinear data dependencies should be applied to existing algorithms to improve them. Brainstorm the idea of hybridity, which consists of different ways to enhance both the technical and adaptive performance.

Interpretability and visualisation can improve high-dimensional data representation and clustering evaluation. Domain-specific applications could optimise these methods for healthcare, finance, and social networks. Scalability and efficiency improvements are also needed, with parallel and distributed techniques being studied to handle enormous datasets. Finally, standardising assessment measures and benchmark datasets would enable fair comparisons between methodologies and improve research repeatability, advancing the field and enabling transformational applications across areas.

REFERENCES

[1] Y. LeCun, C. Cortes, and C. J. Burges, "MNIST handwritten digit database," *ATT Labs*, 2010. [Online]. Available: http://yann.lecun.com/exdb/mnist

[2] Kaggle, "Rice dataset Commeo and Osmancik," 2024. [Online]. Available: https://www.kaggle.com/datasets/muratkokludataset/rice-dataset-commeo-and-osmancik

[3] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, no. 6, pp. 417–441, 1933.

[4] A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Networks*, vol. 13, no. 4–5, pp. 411–430, 2000.

[5] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, Nov. 2008.

[6] S. Dasgupta, "Experiments with random projection," in *Proc. 16th Conf. Uncertainty in Artificial Intelligence (UAI)*, 2000, pp. 143–151.

[7] J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, vol. 1, no. 14, pp. 281–297.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–38, 1977.

[9] Analytics Vidhya, "Top 12 dimensionality reduction techniques," 2018. [Online]. Available: https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-python/

[10] Javatpoint, "Introduction to dimensionality reduction technique," 2024. [Online]. Available: https://www.javatpoint.com/dimensionality-reduction-technique

[11] TensorFlow, "MNIST dataset overview," 2024. [Online]. Available: https://www.tensorflow.org/datasets/catalog/mnist

[12] M. Koklu, "Rice image dataset," Kaggle, 2024. [Online]. Available: https://www.kaggle.com/datasets/muratkokludataset/rice-image-dataset

[13] M. Miyade, "Rice-image-dataset," GitHub, 2019. [Online]. Available: https://github.com/miyade2019/Rice-Image-Dataset

[14] Freecodecamp, "8 clustering algorithms in machine learning that all data scientists should know," 2020. [Online]. Available: https://www.freecodecamp.org/news/8-clustering-algorithms-in-machine-learning-that-all-data-scientists-should-know/

[15] Mendeley Data, "An image dataset of rice varieties," 2024. [Online]. Available: https://data.mendeley.com/datasets/3mn9843tz2/3
[16] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, vol. 15, 2002, pp. 833–840.
[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.
[18] Y. Lin, K. Wang, X. Sun, G. Xu, and X. Wang, "A remote sensing image dataset for cloud removal," *arXiv preprint arXiv:1901.00600*, 2019.
[19] Encord, "Top 12 dimensionality reduction techniques for machine learning," Encord Blog, 2024. [Online]. Available: https://encord.com/blog/dimentionality-reduction-techniques-machine-learning/
[20] Javatpoint, "Clustering in machine learning," 2024. [Online]. Available: https://www.javatpoint.com/clustering-in-machine-learning
[21] Datarundown, "6 different types of clustering: All you need to know," 2024. [Online]. Available: https://datarundown.com/types-of-clustering/
[22] Papers With Code, "RICE dataset," 2024. [Online]. Available: https://paperswithcode.com/dataset/rice
[23] GitHub, "BUPTLdy/RICE_DATASET," 2024. [Online]. Available: https://github.com/BUPTLdy/RICE_DATASET
[24] Kaggle, "MNIST dataset," 2024. [Online]. Available: https://www.kaggle.com/datasets/hojjatk/mnist-dataset
[25] GitHub, "MNIST dataset," 2024. [Online]. Available: https://github.com/cvdfoundation/mnist
[26] Analytics Vidhya, "An introduction to clustering and different methods of clustering," 2016. [Online]. Available: https://www.analyticsvidhya.com/blog/2016/11/an-introduction-to-clustering-and-different-methods-of-clustering/
[27] Medium, "Linear vs non-linear dimensionality reduction: PCA and Kernel PCA," [Online]. Available: https://medium.com/@abhishek8694/linear-vs-non-linear-dimensionality-reduction-pca-and-kernel-pca-10490f345ba9
[28] M. Asif and S. Ullah, "Determinants of support for federalism vs. centralization: A survey of public opinion in Punjab and Khyber Pakhtunkhwa (KP)," *Social Science Review Archives*, vol. 4, no. 1, pp. 2791–2807, 2026, doi: 10.70670/sra.v4i1.1843.
[29] M. Asif and S. Ullah, "Performance voting vs. identity voting: An analysis of electoral behaviour in Pakistani districts," *Journal of Applied Linguistics and TESOL (JALT)*, vol. 9, no. 1, pp. 213–226, 2026, doi: 10.63878/cjssr.v4i1.2079.
[30] M. Asif, A. Ali, and F. A. Shaheen, "Assessing the effects of artificial intelligence in revolutionizing human resource management: A systematic review," *Social Science Review Archives*, vol. 3, no. 4, pp. 2887–2908, 2025, doi: 10.70670/sra.v3i3.1055.
[31] D. Mohiuddin, "Adaptive marketing systems and consumer feedback loops: Implications for market development in emerging economies," *Journal of Business Insight and Innovation*, vol. 5, no. 1, pp. 37–48, 2026.
[32] D. Mohiuddin, "HR tech adoption in digital banking: Implications for workforce development and financial sector growth in emerging economies," *Journal of Business Insight and Innovation*, vol. 4, no. 2, pp. 77–90, 2025.
[33] D. Mohiuddin and D. N. Farhan, "Artificial intelligence in marketing: Ethical challenges and solutions for consumers and society," *Journal of Business Insight and Innovation*, vol. 4, no. 1, pp. 73–87, 2025.
[34] D. Mohiuddin, "Algorithmic hyper-personalization: The double-edged sword of predictive personalization—An empirical investigation," *Journal of Engineering and Computational Intelligence Review*, vol. 2, no. 2, pp. 82–94, 2024.
[35] D. Mohiuddin, "Consumer perceptions and trust in AI-generated advertising: An experimental study in the Pakistani context," *Apex Journal of Social Sciences*, vol. 3, no. 1, pp. 53–68, 2024.